Structuration automatique de flux télévisuels

Camille Guinaudeau

THÈSE INSA Rennes
sous le sceau de l'Université Européenne de Bretagne
pour obtenir le grade de
DOCTEUR DE L'INSA DE RENNES
Spécialité : Informatique

présentée par
Camille Guinaudeau
ÉCOLE DOCTORALE : MATISSE
LABORATOIRE : IRISA/INRIA

Structuration automatique de flux télévisuels

Thèse soutenue le 7 décembre 2011
devant le jury composé de :

Brigitte Grau
Professeur à ENSIIE / *Présidente*

François Yvon
Professeur à l'Université Paris Sud 11 / *Rapporteur*

Philippe Langlais
Professeur à l'Université de Montréal / *Rapporteur*

Patrice Bellot
Professeur à l'Université de Marseille / *Examinateur*

Pascale Sébillot
Professeur à l'INSA de Rennes / *Directrice de thèse*

Guillaume Gravier
Chargé de recherche CNRS / *Co-directeur de thèse*

In theory, there is no difference between theory and practice.
But, in practice, there is.
Jan L. A. van de Snepscheut

Remerciements

Table des matières

Introduction

Contexte de la thèse

Depuis plusieurs années, le média vidéo est devenu prépondérant dans notre façon d'accéder à l'information et au divertissement, prenant peu à peu la place du média textuel. Selon le cabinet d'audit *GfK 7*, chaque famille française possédait en 2007 1,8 téléviseur en moyenne et laissait ce téléviseur allumé environ six heures par jour. Cette modification des comportements a conduit à une augmentation importante du nombre de documents multimédias produits et diffusés chaque année. En France, l'Institut National de l'Audiovisuel (INA), chargé depuis le 1er janvier 1995 de la mise en œuvre du dépôt légal de la radio-télévision, possède des collections multimédias qui comptent plus de 4 millions d'heures de programmes, collections qui s'enrichissent chaque année d'environ 900 000 heures. La mise en place de la Télévision Numérique Terrestre (TNT) et l'ajout de 18 nouvelles chaînes de télévision ont encore augmenté cette quantité, puisque depuis le 1er décembre 2008, ce sont 88 chaînes de télévision et 20 chaînes de radio qui sont collectées 365 jours par an. À ces données professionnelles sont venues s'ajouter récemment les vidéos accessibles sur Internet, faisant exploser le volume de documents multimédias disponibles. À titre d'exemple, 13 millions d'heures de vidéos ont été ajoutées sur le site *YouTube* en 2010.

Afin de rendre exploitables les données archivées par leur institut, les documentalistes de l'INA mettent en place, chaque année, l'analyse et l'indexation de près de 90 000 émissions de télévision et de radio en produisant un résumé et une indexation thématique, sous forme d'une liste de mots clés, de l'émission. Cependant, l'explosion du nombre de vidéos disponibles rend indispensable l'établissement de méthodes automatiques pour l'analyse et la description des flux multimédias. Cet effort de structuration, sans lequel les milliers d'heures de vidéos disponibles resteraient inutilisables, peut prendre différentes formes. L'extraction et la caractérisation de programmes télévisés à partir d'un flux télévisuel, par exemple, permettent de rendre plus accessibles à des utilisateurs les informations contenues dans le flux. Pour cela, il est nécessaire de repérer les frontières de début et de fin de programmes d'une part et de proposer une représentation du contenu des programmes d'autre part. La structuration de données multimédias peut également se situer au niveau des émissions elles-mêmes. Dans ce cadre, l'objectif de la structuration est d'autoriser l'accès des utilisateurs à un point précis de la vidéo. Cela peut correspondre à une action dans une vidéo de rencontre sportive (un but dans un match de football par exemple) ou à un segment de la vidéo dans lequel apparaît une personnalité connue. La structuration fine des émissions doit également fournir aux utilisateurs la possibilité de naviguer au sein d'une collection de segments vidéos traitant d'un sujet similaire et ainsi suivre les évolutions d'un sujet d'actualité. C'est à cet aspect de structuration fine d'émissions de télévision que s'intéresse cette thèse.

Problématique et objectifs de la thèse

Cette thèse présente nos travaux sur la structuration[1] automatique d'émissions télévisuelles. Pour réaliser cette structuration, nous proposons deux approches ayant deux objectifs légèrement différents (*cf.* Figure 1). La première méthode, que nous appellerons *structuration thématique linéaire* dans la suite du document, consiste à organiser une collection de documents de même nature (par exemple une collection de journaux télévisés sur une période de plusieurs jours ou plusieurs semaines) afin d'en faire émerger les principaux thèmes et de permettre aux utilisateurs d'accéder directement aux segments des émissions qui traitent du sujet qui les intéresse. Pour ce faire, nous agissons en deux étapes (*cf.* Figure 1 - partie gauche). Les programmes télévisés sont tout d'abord segmentés thématiquement afin d'extraire des éléments homogènes du point de vue du sujet qu'ils abordent. Ces segments thématiquement homogènes sont ensuite associés les uns aux autres dès lors qu'ils traitent de thématiques similaires. La seconde méthode de structuration que nous avons étudiée, dénommée *structuration thématique hiérarchique* dans la suite du manuscrit, consiste à extraire une structure interne à une émission par le biais d'une segmentation thématique hiérarchique (*cf.* partie droite de la figure 1). Dans ce cas, l'objectif de la méthode est de fournir une organisation détaillée de l'émission, offrant à l'utilisateur la possibilité de *zoomer* sur un sujet précis ou d'avoir une vision plus globale d'un fait d'actualité. Nous choisissons pour ce faire de fonder la segmentation hiérarchique sur une utilisation répétitive d'un algorithme de segmentation linéaire.

La segmentation thématique des émissions de télévision, qu'elle soit linéaire ou hiérarchique, constitue donc le cœur de ce travail de thèse et fait l'objet de la plupart de nos contributions. Les programmes télévisés étant des données multimodales, de nombreux travaux de segmentation thématique ont utilisé des indices vidéos ou audios comme sources d'information sur la structure des documents. Par exemple, (Amaral and Trancoso, 2003) exploite la détection de locuteur afin de repérer le présentateur du journal télévisé, celui-ci introduisant de nouveaux reportages et donc des changements thématiques. Cependant, ces indices sont très dépendants du type de données qu'ils permettent de structurer et ne sont efficaces que dans un cadre précis – la présence d'un présentateur dans l'émission à segmenter est, par exemple, indispensable dans le cadre des travaux de Amaral et Trancoso. Or, l'importante quantité de documents multimédias disponibles n'autorise pas la mise en place de méthodes *ad hoc* à chaque type d'émission et rend nécessaire l'usage de techniques non supervisées capables de segmenter, et de structurer au sens plus large, des émissions tout-venant. Pour combler cette absence de supervision et atteindre notre objectif de structuration de grande quantité de données télévisuelles hétérogènes, nous basons notre travail de segmentation thématique sur la parole prononcée au cours des émissions – celle-ci étant indépendante du type des émissions considérées – et appliquons sur cette parole des méthodes issues du traitement automatique des langues (TAL).

L'un des moyens d'accéder à la parole contenue dans les émissions télévisées est l'utilisation du Télétexte ou du sous-titrage des émissions. Cependant, si les grandes chaînes hertziennes, dont l'audience moyenne annuelle est supérieure à 2,5% de l'audience totale des services de télévision, ont l'obligation légale de sous-titrer l'intégralité des émissions depuis 2010, ce n'est

[1] La structuration est ici prise au sens d'extraction de l'organisation (d'une collection) de documents télévisuels. L'aspect description des programmes n'est que légèrement abordé dans le chapitre 6. Cependant, conscients que la caractérisation du contenu des émissions est essentielle pour accéder à l'information contenue dans le flux, nous revenons sur ce point dans la conclusion en proposant quelques perspectives.

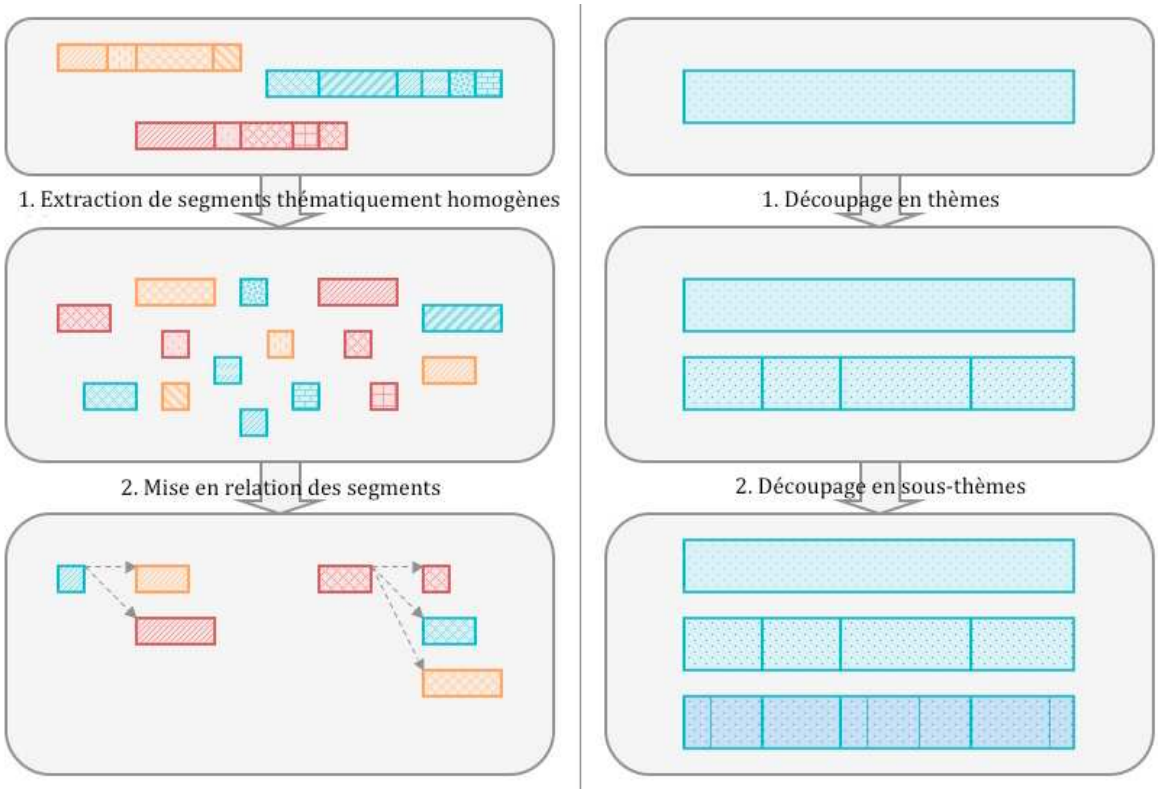


FIG. 1 – Approches pour la structuration thématique linéaire et la structuration thématique hiérarchique

pas le cas des chaînes de la TNT ayant une part d'audience plus faible (comme LCI ou i-Télé)[2]. De plus, si les sous-titres d'émissions diffusées aujourd'hui sont pour la plupart disponibles, ce n'est pas le cas pour des émissions plus anciennes. Afin de pallier ces difficultés, nous avons choisi d'accéder à la parole prononcée durant les émissions par le biais des transcriptions automatiques obtenues grâce à un système de transcription automatique de la parole. Si ces transcriptions permettent de récupérer la parole prononcée dans tout type d'émission, elles possèdent également certaines particularités qui les rendent très différentes du texte écrit.

La technique, développée dans (Utiyama and Isahara, 2001) et fondée sur le critère de cohésion lexicale, qui sert de base à la segmentation thématique mise en place dans cette thèse est malheureusement sensible aux spécificités des transcriptions automatiques qui dégradent fortement ses performances. La première contribution de notre travail consiste donc à proposer une adaptation du critère de cohésion lexicale afin de le rendre plus robuste aux particularités de ces données. Pour ce faire, nous utilisons, de manière conjointe ou indépendante, des connaissances linguistiques et des informations issues de la reconnaissance automatique de la parole et du signal. Ces indices sont obtenus de façon totalement automatique, ce qui distingue notre travail des autres tentatives rencontrées jusqu'à présent dans la littérature. En plus de cet objectif d'adaptation du critère de cohésion lexicale, nous cherchons également à améliorer les performances de l'algorithme de segmentation thématique, en prenant en compte les spécificités de l'oral d'une part et en travaillant sur un perfectionnement plus général de

[2]http ://www.csa.fr/actualite/dossiers/dossiers_detail.php ?id=127347&chap=3235

l'algorithme, c'est-à-dire indépendamment de la qualité des données à traiter – que ce soit des données textuelles classiques ou des données plus particulières telles que les transcriptions automatiques. Pour cela, nous proposons une combinaison de la mesure de la cohésion lexicale avec une détection de ruptures thématiques.

Si la segmentation thématique linéaire constitue le fondement de nos approches de structuration et la phase sur laquelle nous consacrons la plupart de nos efforts, le passage de cette étape à la structuration proprement dite donne également lieu à deux contributions. Premièrement, nous proposons d'étendre des techniques de recherche d'information, utilisées pour la mise en relation de segments thématiquement homogènes, afin de tirer parti des spécificités de nos données orales par le biais de la prosodie. En effet, si les transcriptions obtenues à partir de la parole prononcée dans l'émission nous permettent d'accéder au contenu sémantique de l'émission, la façon dont est prononcée la parole au sein de l'émission est également un vecteur d'information important. Deuxièmement, la segmentation thématique hiérarchique ayant été très peu étudiée jusqu'à présent, nous mettons en place, dans cette thèse, un travail exploratoire examinant différents dispositifs permettant le passage d'une méthode de segmentation thématique linéaire à une technique de segmentation hiérarchique.

Organisation du manuscrit

Ce manuscrit s'organise en 7 chapitres. Les trois premiers sont consacrés au positionnement de nos travaux. Nous présentons tout d'abord en détail nos objectifs de structuration en les comparant aux approches déjà proposées dans la littérature. Dans un deuxième temps, nous décrivons les particularités des transcriptions automatiques ainsi que les difficultés qu'elles entraînent dans notre travail de structuration. Ce deuxième chapitre est également consacré à la présentation des données multimédias qui ont servi de base de test pour nos travaux. Dans le troisième chapitre, nous exposons les indices utilisés tout au long de cette thèse pour adapter les techniques employées pour la structuration, linéaire ou hiérarchique, aux spécificités des transcriptions automatiques. Les chapitres 4 et 5 constituent une deuxième partie dans laquelle nous mettons l'accent sur la segmentation thématique linéaire, qui est la brique de base nécessaire à l'élaboration de nos méthodes de structuration. Le chapitre 4 est consacré à la présentation de la segmentation thématique appliquée à du texte standard, alors que le chapitre 5 présente les méthodes utilisées pour adapter cette segmentation thématique linéaire aux transcriptions automatiques. Finalement, les deux derniers chapitres sont dédiés au passage de la segmentation linéaire à la structuration proprement dite. Dans le chapitre 6, nous décrivons une méthode de structuration *inter-émissions* qui s'appuie sur la mise en relation de segments thématiquement homogènes. Le chapitre 7 est, quant à lui, consacré à la présentation d'un travail plus exploratoire sur la segmentation thématique hiérarchique qui permet de fournir une structuration *intra-émission*. Finalement, nous concluons ce manuscrit en soulignant nos principales contributions et en proposant quelques pistes de recherche ouvertes par nos travaux.

Première partie

Positionnement

Chapitre 1

Structuration automatique de flux TV : état de l'art et positionnement

L'augmentation du nombre de documents multimédias disponibles rend indispensable le développement de méthodes automatiques permettant de structurer et de décrire ces documents afin de faciliter l'accès aux informations qu'ils contiennent. Cette structuration peut prendre différentes formes selon qu'elle cherche à extraire l'architecture interne d'un document ou qu'elle a pour but l'organisation d'une collection. Dans le premier cas, différentes représentations du contenu du document peuvent être envisagées pour permettre un accès rapide à l'information qu'il renferme : résumés automatiques, vidéos ou textuels, tables des matières, *etc.* Dans le second cas, la structuration cherche à positionner les documents les uns par rapport aux autres afin de faciliter la compréhension globale de l'utilisateur par rapport à la collection. Le principal objectif de ce chapitre est de présenter les différents travaux de structuration automatique de documents audiovisuels à travers ces deux motivations que sont la structuration de documents et la structuration de collections. Nous nous positionnons également par rapport à ces travaux en définissant nos objectifs de structuration, dans la section 1.2, ainsi que les approches que nous avons choisi de mettre en œuvre pour y parvenir.

1.1 État de l'art

L'accès à l'information contenue dans des documents multimédias peut s'effectuer de deux façons. L'utilisateur peut chercher à visualiser rapidement un point d'intérêt dans une vidéo particulière, comme un but dans un match de football ou une partie d'une émission abordant un thème particulier. Dans ce cas, il est nécessaire de développer des techniques permettant d'extraire la structure interne des vidéos, techniques décrites dans la section 1.1.1. L'utilisateur peut également souhaiter suivre les évolutions d'un sujet d'actualité au cours du temps, étudier la façon dont différents médias traitent d'un même fait d'actualité ou accéder à toutes les émissions partageant des caractéristiques similaires (la présence des mêmes personnes au sein de la vidéo par exemple). Les techniques de structuration automatique de collections de documents multimédias permettant d'organiser des documents de même nature les uns par rapport aux autres sont présentées dans la partie 1.1.2.

1.1.1 Structuration de documents

Par analogie avec les documents textuels, dont la gestion a été depuis longtemps organisée au sein de bibliothèques et autres moteurs de recherche textuels, l'accès à l'information incluse dans les documents multimédias nécessite l'indexation du contenu de ces documents. Cette indexation peut prendre la forme d'une table des matières ou d'un résumé qui fournissent à l'utilisateur une indication sur la structure d'un document. Cette structuration peut être effectuée à différents niveaux de la vidéo, du plus petit, comme l'image, au plus large tels que les flux télévisuels.

Le niveau de structuration de documents le plus élémentaire est la segmentation en plans. Elle consiste à isoler, dans une vidéo, les séries d'images acquises de manière continue par une caméra et séparées par des transitions de différentes natures, transitions brusques, fondus enchaînés, balayages, *etc.* Les méthodes de segmentation en plans reposent généralement sur la détection de ces transitions par des méthodes fondées sur des mesures de similarités de caractéristiques bas niveau (Lienhart, 2001; Yuan et al., 2007; Smeaton et al., 2010). Ce niveau de structuration étant trop petit pour que le résultat soit sémantiquement interprétable par un utilisateur, la structuration en plans est souvent considérée comme la première étape indispensable à une structuration de plus haut niveau telle que la segmentation en scènes.

La segmentation en scènes, connue aussi sous le nom de macro-segmentation, permet le chapitrage automatique d'une vidéo et offre ainsi la possibilité de déterminer une table des matières du document traité. Sous cette appellation, on retrouve des travaux ayant des objectifs relativement différents, la notion de scène ne possèdant pas de définition claire. La nature d'une scène dépend, en effet, du type ou genre de la vidéo considérée, les scènes étant vues comme un groupe de plans cohérents qui présente un sens pour l'utilisateur. Or, cette cohérence va se baser sur différentes modalités selon que l'on se place dans le cadre d'un film – dans ce cas l'unité de lieu ou de temps va prévaloir – ou dans un journal télévisé dans lequel l'unité thématique sera primordiale. Pour pallier ce problème, la solution envisagée par de nombreux travaux, élaborant ce qu'on appelle des *systèmes spécifiques*, consiste à se restreindre à un cadre d'analyse très contraint permettant d'avoir une définition précise de la notion de scène. Ces *systèmes spécifiques* se concentrent sur des genres de programmes ayant une structure très forte et peu variable. Dans (Kijak, 2003) ou (Delakis, 2006) par exemple, les auteurs cherchent à identifier les phases de jeu dans des matches de tennis. Pour cela, ils utilisent des modèles de Markov cachés pour fusionner des informations provenant à la fois de l'image et de la bande son (Kijak, 2003), éventuellement combinés avec des indices textuels, comme les annonces de points (Delakis, 2006). La structuration de journaux télévisés (JT) a également fait l'objet de nombreuses études, que ce soit par le biais de *systèmes spécifiques*, comme dans (Eickeler and Muller, 1999) où les auteurs proposent le découpage de JT en phases plateaux et reportages en s'appuyant sur des indices vidéos, ou grâce à des techniques de segmentation thématique. La segmentation thématique d'émission autorise en effet la visualisation rapide de la structure du document. Cette structuration peut être effectuée en détectant le présentateur du journal télévisés qui annonce les nouveaux reportages et donc les changements thématiques. Les nombreux travaux qui se sont intéressés à la détection du présentateur peuvent être regroupés en deux grandes familles. Dans la première, les méthodes développées consistent à comparer tous les plans de la vidéo à un modèle, appris préalablement de façon supervisée, représentant les plans dans lesquels apparaît le présentateur (Smoliar et al., 1995). La seconde famille regroupe des travaux mettant en place une détection non supervisée fondée sur des techniques de *clustering* visant à regrouper des plans

ayant un contenu visuel similaire et apparaissant tout au long de la vidéo (Ide et al., 2001; Santo et al., 2006). Cependant ces techniques sont, comme les *systèmes spécifiques*, très dépendantes du types de données à structurer et nécessite la présence d'un présentateur dans l'émission pour fonctionner. Dans (Slaney and Ponceleon, 2001), les auteurs proposent une méthode plus générique ne faisant pas d'hypothèses *a priori* sur la structure de l'émission à segmenter. Dans ce travail, les auteurs combinent une technique employée en traitement du signal, la *scale-space segmentation*, et une méthode de traitement automatique des langues, l'indexation sémantique latente. Cette combinaison, appliquée sur la transcription manuelle d'une émission de *CNN news*, permet d'extraire une structure thématique, et donc une table des matières, du programme traité.

Si la segmentation en scènes constitue un objectif en tant que tel, elle est également un point de départ à la détection d'événements. Elle permet, en effet, de réaliser une présélection des segments de la vidéo qui feront ensuite l'objet d'une analyse plus fine. La détection d'événements qui consiste à repérer, au sein d'un document, des événements ou actions particulières a été principalement appliquée aux retransmissions de rencontres sportives et aux films. Dans les vidéos sportives, elle a pour objectif de retrouver les moments forts d'une rencontre, comme les buts ou les penaltys dans un match de football. Elle permet ainsi de proposer un résumé du match ou d'orienter les utilisateurs vers les segments de la vidéo dans lesquels se déroulent les actions les plus importantes. Ces événements sont détectés grâce à des informations spécifiques au type du programme qui peuvent être sonores, pour repérer les réactions du public (Cabasson and Divakaran, 2003), ou visuelles, comme la détection des ralentis (Zhao et al., 2006). Contrairement aux études conduites pour les rencontres sportives dans lesquelles la détection d'événements consiste à retrouver plusieurs types d'actions, les travaux menés sur les films ont souvent pour objectif de repérer une seule catégorie d'événements. Quelques auteurs se sont, par exemple, intéressés à la détection de scènes violentes. Des indices audios sont couramment employés pour repérer des sons caractéristiques tels que des bruits d'explosions ou d'armes à feu (Moncrieff et al., 2001), éventuellement combinés à des indices visuels comme la présence de sang ou de flammes (Nam et al., 1998).

Alors que les études proposant la mise en place de structuration interne à des programmes télédiffusés sont relativement nombreuses, il existe peu de travaux sur la structuration de très gros documents vidéos, dans laquelle l'unité de structuration est le programme TV. Dans ce cas, l'objectif de la structuration est d'extraire d'un flux télévisuel les différents programmes qu'il contient. Parmi ces travaux, beaucoup reposent sur la détection de répétitions au sein du flux, les répétitions constituant les éléments structurant le flux et séparant les émissions télévisuelles à extraire. Ces éléments structurants sont de natures très diverses, allant des publicités ou des jingles, comme dans (Naturel, 2007) ou (Manson and Berrani, 2010), à des génériques de début et de fin de programmes (Liang et al., 2005). S'il est nécessaire de construire une base de données contenant les éléments répétés à repérer dans (Naturel, 2007), (Manson and Berrani, 2010) propose une technique de détection de séquences répétées, sans connaissance *a priori*, à base de micro-clustering qui ne fait d'hypothèse ni sur la longueur, ni sur la fréquence des séquences à détecter. Ces deux travaux proposent, par ailleurs, une étape finale d'étiquetage qui permet d'indexer les segments du flux non répétés, c'est-à-dire les programmes TV. Cet étiquetage est mis en place grâce à l'alignement entre un guide de programmes électronique (EPG) et les programmes extraits. Les informations contenues dans un guide de programmes sont également utilisées par (Poli, 2007) pour prédire les programmes présents dans le flux. Pour cela, l'auteur se base sur le fait que les chaînes de télévision définissent des grilles de programmes très stables afin de fidéliser le téléspectateur et utilise

une vérité terrain correspondant à une année de diffusion pour entraîner un modèle de Markov caché dans lequel les états représentent les genres des émissions télévisuelles.

1.1.2 Structuration de collections

Les méthodes de structuration présentées dans la sous-section précédente ont pour objectif la mise en évidence de la structure interne des documents afin de proposer aux utilisateurs une table des matières ou un résumé du contenu du document. La structuration de vidéos peut également se faire au niveau de la collection de documents. Dans ce cas, les méthodes développées ont pour but d'organiser la collection afin de rendre son contenu plus facilement accessible à l'utilisateur.

La classification des vidéos est une des techniques de structuration de collections qui consiste à attribuer un ou plusieurs index à un document audiovisuel. La classification en genres est un cas particulier de classification qui cherche à associer un genre ou un sous-genre à une vidéo. Dans (Oger et al., 2010), Oger *et al.* s'intéressent à la classification des vidéos en six genres : clips de musique, publicité, dessins animés, documentaires, journaux télévisés, sport et films. Pour cela, les auteurs utilisent les transcriptions automatiques de la parole prononcée dans les documents afin de prendre en compte les particularités stylistiques de chaque genre. Ils s'intéressent également à des indices ayant prouvé leur efficacité dans des tâches de caractérisation de genre de textes écrits, telles que les séquences de mots les plus fréquentes. Dans le projet CoP (*Content Processing*, (Fischer et al., 1995)), Fischer *et al.* utilisent une méthode de détection de coupures et de mouvements dans la vidéo ainsi que des informations statistiques sur les couleurs afin de classer différentes vidéos en quatre genres : informations, sport, publicité et dessins animés. Ils prennent également en considération des statistiques audios, telles que la fréquence et l'amplitude du signal. Dans cet article, Fisher *et al.* analysent les caractéristiques des films (mouvements, longueur des scènes, *etc.*) pour leur associer un genre et cherchent ensuite à faire correspondre ce genre avec ceux appris automatiquement sur un grand nombre de données. Roach *et al.* (Roach et al., 2001) utilisent, quant à eux, un classifieur statistique pour catégoriser les émissions à partir d'indices à la fois vidéos et audios qui sont dans un premier temps employés séparément puis combinés linéairement. Si la classification d'émissions en genres se fonde sur un ensemble défini de classes qui est habituellement fermé, Roach *et al.* introduisent dans leur étude une nouvelle classe à laquelle appartiennent les émissions dont le genre n'est pas spécifié. Leur travail est, de ce fait, plus proche d'une vérification de l'appartenance d'une vidéo à un genre particulier que d'une classification proprement dite. Même si les études semblent privilégier la combinaison d'indices, qu'ils soient vidéos et audios ou audios et textuels, certains travaux se basent sur un seul type d'indices, comme (Liu et al., 1998) qui opère une classification des programmes télévisuels en n'étudiant que les indices audios. Liu *et al.* utilisent, ainsi, huit propriétés au niveau des trames et quatorze propriétés au niveau du clip vidéo. Ces propriétés sont extraites afin d'entraîner les modèle de Markov cachés ergodiques utilisés pour classer les émissions en cinq genres distincts.

Si la classification en genres des émissions permet de regrouper les documents qui partagent une caractéristique stylistique commune, certains travaux se sont penchés sur le regroupement de documents vidéos abordant les mêmes thématiques. Ces études utilisent la parole prononcée dans les vidéos comme indice de similarité entre deux vidéos. Un travail représentatif de ce qui se fait en parole peut être trouvé dans (Yang et al., 1999). Dans cet article, Yang *et al.*

représentent les documents par des vecteurs de mots clés, pondérés grâce à un poids *tf-idf*[1]. Une mesure cosinus est ensuite utilisée pour calculer la similarité entre les paires de documents et une approche k-NN effectue la classification proprement dite. En plus de vecteurs de mots clés pondérés, (Hsu and Chang, 2006) se base sur des indices visuels bas niveau, tels que la détection de *visual near-duplicates* et sur des concepts sémantiques haut niveau extraits automatiquement à partir de la vidéo. Ces concepts visuels sont ceux définis dans le cadre de la campagne d'évaluation TRECVID (Smeaton et al., 2006) et peuvent être de natures très différentes (visage, ciel, personne qui marche, scène extérieure, *etc.*). Ces indices sont combinés linéairement et une approche k-NN est utilisée pour estimer la pertinence d'un document vidéo par rapport à un thème.

Le regroupement de vidéos présentant des similarités thématiques a également fait l'objet de nombreux travaux cherchant à reconnaître et associer les documents abordant les mêmes faits d'actualité. Ces travaux sur la détection et le suivi d'événements sont devenus très populaires grâce au projet *Topic Detection and Tracking* (TDT) lancé en 1997. Ce projet consiste à élaborer des méthodes d'extraction de structure thématique à partir de flux d'actualités fournis par différents médias et dans différentes langues (Allan, 2002a). Les méthodes développées cherchent à catégoriser les nouveaux documents d'actualité qui apparaissent chaque jour et qui peuvent être soit le développement d'une histoire déjà relatée, soit l'introduction d'un nouveau sujet. Dans (Wu et al., 2006), les auteurs utilisent des indices textuels, extraits de la transcription de la bande sonore, combinés à des indices vidéos. Ils proposent une technique de *co-clustering* associée à un calcul de similarité afin de détecter la redondance ou la nouveauté d'une vidéo par rapport à celles déjà traitées. Les vidéos redondantes sont par la suite écartées afin de proposer aux utilisateurs une vue plus synthétique du sujet considéré. Avec le même objectif de simplification, les auteurs de (Ide et al., 2005) se basent sur les relations entre les différents protagonistes apparaissant dans les vidéos pour structurer un flux de faits d'actualité mis à jour chaque nuit.

1.2 Positionnement

Dans cette section, nous présentons la démarche originale que nous avons mis en place pour la structuration automatique d'émissions de télévision. Nous décrivons tout d'abord les deux procédures de structuration développées, visant deux objectifs légèrement différents. Puis, nous exposons une vue détaillée de notre approche, fondée sur l'utilisation des transcriptions automatiques de la parole contenue dans les émissions.

1.2.1 Objectifs de structuration

Comme nous l'avons souligné plus tôt dans ce chapitre, la structuration automatique de documents multimédias peut prendre différentes formes selon qu'elle s'attache à structurer des documents ou des collections de documents. Au cours de ce travail de thèse, nous proposons deux techniques de structuration, la *structuration thématique linéaire* et la *structuration thématique hiérarchique*.

La première approche a été développée dans le but de permettre à des utilisateurs de naviguer au sein d'une collection de documents de même nature et se rapproche, de ce fait, des techniques de *Topic Detection and Tracking* décrites dans la partie 1.1.2. Cette *structuration thématique linéaire*, qui consiste à mettre en relation des segments thématiquement

[1]La méthode de pondération *tf-idf* est très couramment utilisée un recherche d'information et sera décrite plus en détail dans le chapitre 6.

homogènes abordant des thématiques similaires, se divise en deux étapes. Premièrement, des segments traitant d'un unique sujet sont extraits des émissions par le biais d'un algorithme de segmentation thématique linéaire. Puis, la similarité thématique de ces segments est évaluée grâce à une technique issue de la recherche d'information. Dans ce cadre, chaque segment est représenté par un vecteur caractéristique de son contenu puis une mesure de similarité est calculée entre les différents vecteurs, deux segments étant considérés comme abordant une même thématique lorsque la mesure de similarité dépasse un certain seuil.

La seconde méthode de structuration que nous avons étudiée, dénommée *structuration thématique hiérarchique*, consiste à extraire une structure interne à une émission par le biais d'une segmentation thématique hiérarchique. Dans ce cas, l'objectif de la méthode, similaire à celui des techniques de segmentation en scènes présentées dans la partie 1.1.1, est de fournir une organisation détaillée de l'émission, offrant à l'utilisateur la possibilité de *zoomer* sur un sujet précis ou d'avoir une vision plus globale d'un fait d'actualité. La segmentation hiérarchique ayant été très peu étudiée jusqu'à présent, nous proposons dans cette thèse un travail exploratoire examinant différents dispositifs permettant le passage d'une méthode de *segmentation thématique linéaire* à une technique de *segmentation thématique hiérarchique*. Nous choisissons pour ce faire de fonder la segmentation hiérarchique sur une utilisation répétitive d'un algorithme de segmentation linéaire, c'est-à-dire en re-appliquant l'algorithme de segmentation linéaire sur des segments thématiquement homogènes obtenus lors d'une première exécution de l'algorithme.

1.2.2 Approche retenue

Les méthodes de structuration automatique de documents audiovisuels présentées dans la section 1.1 exploitent indifféremment les indices disponibles au sein des vidéos, qu'ils soient audios, vidéos ou textuels. Notre objectif de structuration automatique de programmes télévisuels se démarque des études précédentes par notre volonté de développer des méthodes génériques permettant de traiter différents types d'émissions. Cette philosophie nous imposant de fonder nos techniques de structuration sur des indices indépendants du genre des documents traités, nous avons choisi de nous appuyer sur la parole prononcée dans les émissions. Cette parole, accessible par le biais des transcriptions automatiques fournies par un système de reconnaissance automatique de la parole (RAP), présente en effet l'avantage, en plus de permettre un accès au contenu sémantique des émissions, d'apparaître dans tous types de programmes télévisés : documentaires, sport, journaux télévisés, *talk shows*, météo, *etc.*

La structuration des émissions, qu'elle soit linéaire ou hiérarchique, est ainsi mise en place en appliquant des méthodes issues du traitement automatique des langues et de la recherche d'information sur les transcriptions automatiques. Cependant, l'utilisation de méthodes développées pour du texte écrit sur des transcriptions n'est pas immédiate, ces dernières possédant certaines particularités. Premièrement, ces données ne contiennent ni ponctuation ni majuscule ; elles ne sont donc pas structurées en phrases comme un texte classique mais en unités appelées groupes de souffle, qui correspondent à la parole prononcée par un locuteur entre deux respirations. De plus, le taux d'erreurs de notre système de RAP, même s'il reste raisonnable sur des émissions comme les journaux télévisés, peut atteindre 70% pour des émissions telles que des films ou des *talk shows*[2], rendant impossible l'utilisation de certains indices tels que les marqueurs discursifs. En réponse à ces contraintes, nous utilisons des méthodes fondées

[2]Le système de reconnaissance automatique de la parole utilisé dans ce travail est présenté dans le chapitre 2. Les caractéristiques des transcriptions automatiques ainsi que nos corpora y sont également décrits.

sur le critère de la cohésion lexicale qui repose principalement sur la répétition de vocabulaire. Ce critère, couramment utilisé en traitement automatique des langues et en recherche d'information, fournit de bons résultats mais reste cependant sensible aux particularités des transcriptions. Les erreurs de transcription vont, en effet, venir perturber le calcul de la cohésion lexicale, un mot qui se répète ne pouvant être détecté comme tel si l'une de ces occurrences n'est pas reconnue correctement par le système de RAP. De plus, les émissions télévisuelles peuvent contenir peu de répétitions de vocabulaire – elles sont, en effet, souvent caractérisées par un emploi massif de synonymes – ce qui pénalise l'estimation de la cohésion lexicale.

Un des objectifs de cette thèse consiste à proposer différentes techniques afin d'adapter le critère de la cohésion lexicale aux spécificités des transcriptions automatiques de vidéos professionnelles. Pour surmonter les difficultés liées aux particularités des transcriptions, certains travaux ont suggéré d'ajouter des indices propres aux documents oraux au seul critère de cohésion lexicale. De nombreuses études ((Grosz and Sidner, 1986; Litman and Passonneau, 1995) et (Beeferman et al., 1997) entre autres) exploitent des marqueurs de discours (tels que *now* ou *by the way*) pour améliorer la détection de frontières thématiques dans des documents oraux. Cependant, ces indices sont très dépendants du type de documents considérés et ne conviennent pas à notre volonté de définir des méthodes de structuration suffisamment génériques pour traiter différentes catégories d'émissions.

L'adaptation proposée dans ce travail va donc être principalement mise en œuvre grâce à l'intégration de sources d'informations additionnelles indépendantes du type des documents étudiés et extraites de façon non supervisée. Certains de ces indices complémentaires, décrits dans le chapitre 3, ont pour but de rendre le critère de la cohésion lexicale robuste aux erreurs de transcription ainsi qu'à la faible répétition de vocabulaire dans nos documents. Nous utilisons, dans ce cas, des mesures de confiance, fournies par le système de RAP et traduisant la probabilité qu'un mot soit correctement transcrit, et des relations sémantiques. L'ajout d'information a également pour objectif de tirer parti des différentes modalités qui composent nos documents audiovisuels. La prise en compte de la prosodie nous autorise ainsi à nous intéresser, non seulement au contenu sémantique de nos émissions, mais également à la façon dont est prononcée la parole au sein des programmes.

Ces indices vont être employés, dans un premier temps, dans le cadre de la segmentation thématique de nos émissions télévisuelles. Les deux approches de structuration de documents audiovisuels développées dans cette thèse reposant sur une étape préalable de segmentation thématique linéaire, celle-ci va, en effet, faire l'objet de la plupart de nos contributions. Premièrement, le calcul de la cohésion lexicale sur lequel repose l'algorithme de segmentation thématique de Utiyama et Isahara (Utiyama and Isahara, 2001), qui sert de base à nos méthodes, est modifié afin de : donner moins de poids aux mots potentiellement incorrects, mettre en évidence les liens sémantiques existant entre les mots des segments thématiques et favoriser les mots proéminents dans le discours des journalistes. Pour adapter l'algorithme de segmentation thématique aux spécificités de nos données, nous proposons également d'utiliser des techniques d'interpolation de modèles de langue afin de mieux estimer la cohésion lexicale des segments courts, celle-ci étant calculée par le biais d'une probabilité généralisée calculée à partir d'un modèle de langue *unigramme*. Finalement, notre dernière contribution concernant la segmentation thématique consiste à prendre en compte à la fois la *mesure de la cohésion lexicale*, telle que définie dans (Utiyama and Isahara, 2001), et la *détection de rupture de la cohésion lexicale*, couramment utilisée en segmentation thématique par des méthodes de segmentation que nous appellons *locales* (Hearst, 1997; Ferret et al., 1998; Claveau and Lefèvre, 2011a). Cette optimisation, présentée dans le chapitre 4, permet de fusionner dans un seul

algorithme deux tendances habituellement utilisées indépendamment l'une de l'autre.

À partir de la segmentation thématique des émissions télévisuelles, nous proposons, dans cette thèse, deux approches différentes pour atteindre nos objectifs de structuration. Dans un premier temps, nous mettons en relation les segments thématiquement homogènes résultant de la phase de segmentation afin d'obtenir une structuration thématique linéaire grâce à laquelle les utilisateurs pourront suivre les évolutions d'un sujet d'actualité. Afin de mettre en place cette structuration thématique linéaire, présentée dans le chapitre 6, nous utilisons une méthode classique de recherche d'information basée sur une représentation vectorielle des segments thématiques. Pour représenter les segments thématiques, les méthodes de recherche d'information sélectionnent, dans le document, les mots qu'elles jugent les plus caractéristiques de son contenu. La sélection de ces mots est généralement fondée sur la fréquence d'apparition des mots dans le document. Or, dans les documents audiovisuels, la façon dont un mot est prononcé constitue un vecteur d'information important concernant la valeur informative de ce mot. Il a, en effet, été montré que, dans le cadre de recherche d'information dans des documents oraux de langue chinoise, l'utilisation de la prosodie permettait d'améliorer les performances du système (Chen et al., 2001). De telles études n'ayant jamais été effectuées, à notre connaissance, sur la langue française, nous proposons, dans ce travail, une évaluation des capacités de la prosodie à aider des méthodes de recherche d'information dans des documents oraux français. Nous utilisons, pour cela, les indices prosodiques, employés précédemment pour la segmentation thématique, afin de favoriser les mots proéminents. De plus, les segments abordant des thématiques similaires n'utilisant pas toujours un vocabulaire commun, nous intégrons des relations sémantiques dans le calcul de la similarité entre les vecteurs caractéristiques des segments. Ce chapitre 6 est également l'occasion de démontrer le bon fonctionnement de notre méthode grâce à deux applications, utilisées pour la structuration de journaux télévisés.

Finalement, la seconde approche de structuration d'émissions télévisuelles, qui consiste à développer une structuration hiérarchique des émissions, permet aux utilisateurs d'accéder d'un seul coup d'œil à l'organisation détaillée du programme en question. Pour réaliser cette structuration thématique hiérarchique, nous avons souhaiter travailler sur la segmentation hiérarchique des émissions. Le travail présenté dans ce cadre au chapitre 7 est plus exploratoire, la notion de segmentation thématique hiérarchique ayant été très peu étudiée. Nous décrivons dans ce manuscrit, les différentes pistes examinées, fondées sur une utilisation répétitive d'un algorithme de segmentation linéaire.

1.3 Bilan du chapitre

Dans ce chapitre, nous avons présenté un état de l'art des méthodes développées pour la structuration automatique de documents télévisuels. Ces travaux, qu'ils cherchent à extraire l'organisation interne d'un document ou à structurer une collection d'émissions, sont généralement spécifiques à un genre d'émissions particulier. À la lumière de cet état de l'art, nous avons proposé deux approches de structuration, une *structuration thématique linéaire* et une *structuration thématique hiérarchique*, suffisamment génériques pour être applicables sur différents types d'émissions. Pour cela, les techniques développées dans cette thèse reposent sur l'utilisation du critère de la cohésion lexicale calculé sur les transcriptions automatiques de la parole contenue dans les émissions.

Chapitre 2

Transcriptions automatiques de programmes TV

Afin de développer des méthodes de structuration automatiques suffisamment génériques pour être utilisées sur différents types d'émissions, nous appliquons, dans cette thèse, des méthodes issues du traitement automatique des langues et de la recherche d'information sur les transcriptions automatiques de la parole contenue dans nos programmes télévisuels. Si ces données nous permettent de satisfaire notre contrainte de généricité, elles possèdent également certaines particularités. La première section de ce chapitre est consacrée à la présentation du fonctionnement général d'un système de reconnaissance automatique de la parole, ainsi qu'à la description détaillée du système IRENE employé pour transcrire nos données télévisuelles. Si cette première section nous permet d'exposer certaines particularités inhérentes aux transcriptions automatiques, nous souhaitons présenter dans une seconde partie les spécificités de nos données, liées à l'origine professionnelle des vidéos transcrites. Cette seconde section nous offre, également, l'occasion de décrire en détail les corpora utilisés comme base de test pour nos méthodes de structuration automatique.

2.1 Système de reconnaissance automatique de la parole

Les transcriptions automatiques de la parole constituant l'information de base sur laquelle vont reposer toutes nos techniques de structuration, il est essentiel d'avoir une bonne vision de la façon dont ces transcriptions sont obtenues ainsi que de leurs particularités. Dans cette section, nous décrivons brièvement le principe de fonctionnement des systèmes de transcription automatique de la parole (pour plus de détail, voir (Rabiner and Juang, 1993; Jelinek, 1998; Haton et al., 2006)) ainsi que les sorties proposées par ces systèmes. Puis, nous présentons plus en détail le système IRENE employé pour transcrire nos programmes télévisuels.

2.1.1 Principe

Le but d'un système de reconnaissance est de fournir la transcription textuelle de la parole contenue dans un signal d'entrée audio. Dans le cadre d'une modélisation statistique de la parole, cette tâche équivaut à rechercher parmi l'ensemble des séquences de mots possibles à partir d'un vocabulaire fixé, la séquence la plus probable étant donnée une séquence de caractéristiques acoustiques observées à partir du signal d'entrée.

Un système de transcription automatique de la parole est composé de trois éléments fondamentaux : un modèle de langue qui évalue la probabilité d'une séquence de mots, un modèle acoustique qui calcule la vraisemblance du signal sachant une séquence de mots W, et un lexique phonétisé qui permet de faire le lien entre les représentations sur lesquelles se basent ces deux modèles. Ce lexique permet d'associer à chaque mot du vocabulaire du système de reconnaissance une liste de prononciations possibles, représentées sous la forme de séquence de phonèmes[1].

Le vocabulaire, définissant l'ensemble des mots qui peuvent être reconnus, constitue l'un des éléments déterminants du système de reconnaissance automatique de la parole. Il limite, en effet, les sorties du système aux seuls mots qu'il contient, entraînant des erreurs de transcription dès lors qu'un mot prononcé dans le signal d'entrée n'appartient pas à ce vocabulaire ; on parle alors de mots *hors vocabulaire*. Le modèle de langue, qui permet de calculer la probabilité *a priori* d'une séquence de mots W, est un autre composant primordial d'un système de transcription automatique de la parole. Dans le cadre de modèles de langue statistiques, la probabilité *a priori* d'une séquence de mots W est décomposée en probabilités conditionnelles où la probabilité de chaque mot w_i de W est calculée sachant l'historique des mots $w_1 ... w_{i-1}$ qui le précédent. Chacune de ces probabilités conditionnelles est estimée à partir d'un vaste corpus textuel d'apprentissage. Cependant, lorsque le vocabulaire de ce corpus d'apprentissage est trop grand, l'estimation de ces probabilités devient impossible. Pour pallier ce problème, la stratégie la plus répandue consiste à approximer l'historique en ne conservant que les n mots les plus récents. Dans ce cas, on parle de modèle de langue n-grammes, n étant généralement compris entre 2 et 5.

2.1.2 Sorties

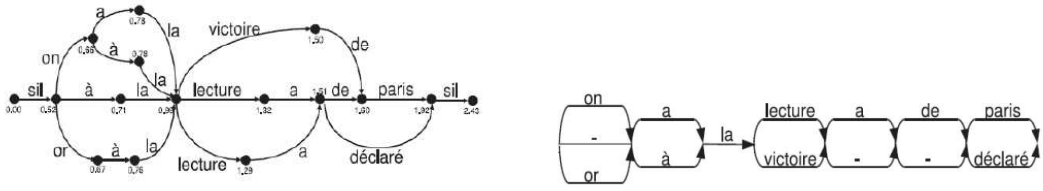
Un système de reconnaissance automatique de la parole peut fournir différents types de sortie. En plus de la transcription finale ou des meilleures hypothèses de transcription retrouvées, le système propose, en effet, des sorties intermédiaires ainsi que des informations, calculées *a posteriori*, sur la qualité des transcriptions : les mesures de confiance.

Hypothèses de transcription La sortie par défaut d'un système de reconnaissance est la transcription textuelle de la parole, qui correspond à la séquence de mots considérée comme la plus probable par le système de transcription. Les transcriptions automatiques présentent plusieurs caractéristiques qui les différencient du texte écrit. Premièrement, ces données ne contiennent ni ponctuation ni, dans la plupart des systèmes, majuscule ; elles ne sont donc pas structurées en phrases comme un texte classique mais en unités appelées groupes de souffle, qui correspondent à la parole prononcée par un locuteur entre deux respirations. De plus, les transcriptions contiennent un nombre potentiellement important de mots mal transcrits. Ces erreurs de transcription peuvent être liées à la qualité de l'enregistrement – enregistrement studio ou extérieur –, à la présence ou non de bruits de fond, d'applaudissements, à la différence de styles de parole ou à la présence de mots hors vocabulaire. La qualité d'une transcription est évaluée grâce à une mesure appelée taux d'erreur mot (ou *WER* pour *word error rate*) correspondant à la distance minimale d'édition entre cette transcription et une transcription de référence, rapportée au nombre de mots de la référence. La figure 2.1 présente un exemple de transcription automatique, associée, à gauche, à sa transcription de référence.

[1] Les phonèmes sont les unités représentant les sons élémentaires d'une langue.

Transcription manuelle	**Transcription automatique**
Dix-neuf cent quatre vingt-deux, un évènement vient de se produire il s'appelle Amandine. Trois kilos quatre, cinquante et un centimètres, le premier bébé éprouvette français est né. Ici, le bébé exploit qui a un an soufflera ce mois-ci ses vingt-cinq bougies.	dix neuf cent quatre-vingt-deux un événement vient de se produire il s'appelle *aman dina* trois kilos quatre cinquante-et-un centimètres le premier bébé *éprouvait* français est né ici le bébé *exploite y a* un an soufflera ce mois ci ses vingt-cinq bougies

FIG. 2.1 – Transcription manuelle et automatique extraite du journal télévisé de France 2 diffusé le 7 février 2007. Les mots en italique correspondent à des erreurs de transcription.



Graphe de mots Réseau de confusion

FIG. 2.2 – Sorties intermédiaires d'un système de transcription automatique de la parole.

Si la sortie de base d'un système de reconnaissance correspond à la meilleure séquence trouvée, le système peut fournir également la liste ordonnée des N meilleures hypothèses de transcription.

Graphe de mots et réseaux de confusion Outre les sorties finales que sont les hypothèses de transcription, un système de reconnaissance propose des sorties intermédiaires – les graphes de mots et les réseaux de confusion – qui permettent de visualiser les différentes hypothèses de mots considérées lors de la production de la transcription textuelle. Les graphes de mots correspondent, pour un groupe de souffle donné, à un graphe dans lequel les nœuds sont des instants du signal et où les arcs représentent des hypothèses de mots accompagnés de leur vraisemblance acoustique et de leur probabilité linguistique. Ces graphes pouvant atteindre des tailles importantes, il peut être intéressant de les compacter sous la forme de réseaux de confusion. Ces réseaux peuvent être vus comme des graphes de mots dont certains nœuds ont été fusionnés en alignant temporellement les meilleurs hypothèses issues des graphes de mots (*cf.* Figure 2.2). Dans ce cas, les arcs, pondérés par la probabilité *a posteriori* de chaque mot w, représentent les hypothèses des mots en concurrence pour chaque section temporelle du signal.

Mesures de confiance Afin de traduire la confiance que le système de reconnaissance automatique de la parole porte à la transcription produite, un score, compris entre 0 et 1, est associé à chacun des mots transcrits. Ces mesures de confiance traduisent une estimation de la probabilité pour un mot d'être correctement reconnu par le système, c'est-à-dire que plus la valeur est proche de 1 plus il est probable que la transcription du mot soit fiable. Elles sont généralement calculées à partir des probabilités *a posteriori* ou d'informations dérivées des graphes de mots. Si ces mesures de confiance permettent de prendre en compte d'éventuelles

erreurs de transcription, il est à noter qu'elles ne sont elle-mêmes pas toujours fiables et doivent donc être considérées avec prudence.

2.1.3 Le système IRENE

Le système IRENE, utilisé dans ce travail pour transcrire nos données télévisuelles, a été développé originellement pour la transcription de journaux radiophoniques en langue française. Il repose sur les composants suivants : un vocabulaire de 65 000 mots pour un total de 80 000 prononciations dans le lexique phonétisé ; deux modèles de langue fondés sur ce vocabulaire, un modèle trigramme et un modèle quadrigramme, ainsi que deux modèles acoustiques prenant en compte, ou non, le contexte phonétique des mots prononcés. L'obtention d'une transcription textuelle est mise en œuvre grâce à une stratégie multi passes permettant de réduire progressivement l'ensemble des transcriptions candidates et d'employer des modèles, acoustiques et linguistiques, de plus en plus complexes. Dans sa phase finale, le système IRENE utilise le modèle de langue quadigramme et le modèle acoustique prenant en compte le contexte phonétique pour produire des listes de 1 000 hypothèses de transcription. Ces hypothèses sont ensuite ré-ordonnées, de manière à favoriser les séquences de mots les plus cohérentes sur le plan morphosyntaxique, afin de produire la transcription textuelle finale (Huet et al., 2010).

Les modèles acoustiques sont appris sur environ 250 heures de journaux radiophoniques appartenant aux données françaises de la campagne ESTER 2 (Galliano et al., 2009). Les probabilités des modèles de langue sont, quant à elles, estimées sur 500 millions de mots extraits d'articles des journaux *Le Monde* et *l'Humanité* entre 1987 et 2002, et interpolées avec des probabilités d'un modèle de langue appris sur plus de 2 millions de mots correspondant aux transcriptions de référence de journaux radiophoniques. Le système IRENE présente un taux d'erreur de transcription de 16% sur les données non accentuées des journaux de la campagne ESTER 2.

Les mesures de confiance, proposées par le système, sont générées grâce à la combinaison de scores acoustiques, linguistiques et morphosyntaxiques (Huet et al., 2010).

2.2 Transcriptions automatiques de programmes TV

Si, comme nous l'avons vu dans la section précédente, les transcriptions automatiques possèdent des caractéristiques intrinsèques les différenciant du texte écrit, l'origine de la parole transcrite impacte également sur la qualité des transcriptions. Nous souhaitons présenter dans cette section les particularités, liées à l'origine professionnelle de nos données, des transcriptions que nous manipulons au cours de cette thèse. Nous décrivons également, dans la section 2.2.2, le détail des trois corpora utilisés pour évaluer les performances de nos méthodes de structuration automatique.

2.2.1 Particularités de transcriptions de programmes TV

Outre les particularités inhérentes aux transcriptions automatiques – absence de ponctuation et de majuscule, erreurs de transcription – les transcriptions automatiques de programmes télévisuels possèdent certaines spécificités. D'une part, les données télévisuelles sont caractérisées par une variation importante de la qualité du signal, ou de la parole prononcée, entraînant une dégradation importante de la qualité des transcriptions. En effet, contrairement à des données radiophoniques, où la parole doit être nécessairement compréhensible par les auditeurs,

une mauvaise réception de la parole, dans des documents audiovisuels, peut être compensée par l'ajout de sous-titres. Il n'est, en effet, pas rare de trouver dans des émissions télévisuelles des cas où l'enregistrement des vidéos s'effectue dans des environnements bruyants rendant inaudible la parole prononcée. De même, certaines personnes interrogées dans le cadre de programmes télévisés peuvent s'exprimer dans une langue étrangère ou avec un fort accent, perturbant ainsi la reconnaissance automatique de la parole. D'autre part, une grande partie de la parole prononcée dans le cadre d'émissions de télévision professionnelle correspond à de la parole préparée. Or, ce type de parole pénalise fortement le critère de cohésion lexicale sur lequel se fondent nos méthodes de structuration. La parole préparée est, en effet, caractérisée par l'emploi massif de synonymes, notamment dans les émissions d'actualités dans lesquelles les présentateurs cherchent à éviter les répétitions de vocabulaire.

2.2.2 Description des corpora

Afin d'évaluer les performances des méthodes de structuration automatique développées, trois corpora ont été utilisés comme bases de test. Le premier corpus, noté corpus *JT* dans la suite du manuscrit, est composé de 56 journaux télévisés diffusés en février et mars 2007 sur la chaîne de télévision France 2. Les deux autres corpora, d'une durée de 14 heures chacun, sont constitués d'émissions de reportages : *Sept à Huit* pour le premier et *Envoyé Spécial* pour le second. Le corpus *Sept à Huit* contient 16 émissions diffusées en 2008 et 2009 sur TF1 et le corpus *Envoyé Spécial* compte 7 émissions programmées ces mêmes années sur France 2. Comme le montre le tableau 2.1, ces trois corpora possèdent des caractéristiques différentes. Premièrement, la durée des émissions est relativement variable d'un type d'émission à l'autre, de 40 minutes pour les journaux télévisés à 2h pour les émissions *Envoyé Spécial*. De plus, la taille des segments thématiques composant les différentes émissions, et obtenus grâce à une segmentation manuelle de référence, varie de 1,6 minute en moyenne pour les JT à 55 minutes. Ces tailles de segments sont par ailleurs très variables au sein d'une même émission pour les corpora *JT* et *Sept à Huit* puisque les segments les plus courts ne durent que quelques secondes contre plus d'une dizaine de minutes pour les plus longs[2]. Les transcriptions automatiques de ces corpora présentent également des dissimilarités tant sur le nombre de mots pleins présents dans chaque segment que sur le nombre de répétitions. Dans le corpus de journaux télévisés, seuls 6,3 mots pleins se répètent, en moyenne, dans chaque segment, ce faible taux de répétition s'expliquant, d'une part, par la faible taille des segments thématiques et, d'autre part, par l'usage de synonymes comme nous l'avons signalé dans la section précédente. Finalement, le taux d'erreur des transcriptions diffère en fonction du type d'émission. Les journaux télévisés sont les mieux transcrits – le corpus est en effet associé à un taux d'erreur de VALEUR –, la proportion d'enregistrement en studio étant plus importante dans ce corpus. Les émissions de reportages, privilégiant les investigations sur le terrain, ont un taux d'erreur plus élevé, VALEUR pour les émissions *Sept à Huit* et VALEUR pour le corpus *Envoyé Spécial*. Remarque:`Mettre à jour ce paragraphe en fonction des taux d'erreur calculés`.

Les deux premiers corpora, composés des journaux télévisés et des émissions de reportages *Sept à Huit*, ont été employés pour tester la structuration thématique linéaire, c'est-à-dire la segmentation thématique linéaire suivie de la mise en relation de segments thématiquement homogènes. Le troisième corpus, constitué des émissions *Envoyé Spécial*, a, quant à lui, été

[2]Dans le corpus de journaux télévisés, ces segments très courts correspondent à des brèves. Pour les émissions *Sept à Huit*, ces sections thématiques de quelques secondes sont des annonces faites par le présentateur, exposant les sujets qui seront abordés dans la suite de l'émission.

TAB. 2.1 – Description des corpora

	durée du corpus	nb de documents	nb de segments	durée moyenne des segments	nb moyen de mots répétés par segment	nb de mots dans chaque segment	taux d'erreur mot
JT	32 h	56	1203	1,6 min *max* : 11 min *min* : 3,4 sec	6.3	107	
Sept à Huit	14 h	16	86	8,6 min *max* : 21 min *min* : 5,5 sec	38	424	
Envoyé Spécial	14 h	7	26	33 min *max* : 55 min *min* : 22 min	142	1639	

plus spécifiquement consacré à l'évaluation de la structuration thématique hiérarchique. Les performances de nos deux méthodes de structuration n'ont pas été estimées sur les mêmes corpora pour deux raisons. Premièrement, contrairement aux corpora *JT* et *Sept à Huit,* le corpus *Envoyé Spécial* étant composé de segments thématiques contenant une répétition de vocabulaire importante, les méthodes de segmentation thématique développées pour du texte sont tout à fait applicables sur les transcriptions de ce corpus et ne nécessitent pas de phase d'adaptation[3]. De plus, notre travail sur la segmentation thématique hiérarchique n'étant encore qu'exploratoire, nous avons choisi de tester notre technique de segmentation thématique hiérarchique uniquement sur des programmes télévisuels présentant une structure thématique hiérarchique claire, c'est-à-dire les émissions *Envoyé Spécial*.

2.3 Bilan du chapitre

Dans ce chapitre, nous avons présenté le système de reconnaissance automatique de la parole employé dans cette thèse pour transcrire nos émissions télévisuelles. Nous avons également détaillé les particularités de ces transcriptions, erreurs de transcriptions et faible répétition du vocabulaire, qui vont avoir un impact négatif sur le critère de cohésion lexicale sur lequel repose nos méthodes de structuration. Dans le chapitre suivant, nous décrivons les indices, ainsi que leur méthodes d'extraction, utilisés pour rendre la cohésion lexicale robuste à ces spécificités.

[3]L'algorithme développé par Utiyama et Isahara fournit des valeurs de rappel et de précision égales à 1 pour 6 des 7 émissions composant le corpus.

Chapitre 3

Indices utiles à l'adaptation de la cohésion lexicale pour la structuration de documents audiovisuels

La notion de cohésion lexicale est couramment utilisée dans de nombreux travaux de traitement automatique des langues. Ce critère, faisant référence aux relations lexicales qui existent au sein d'un texte et lui donne une certaine unité, est mis en œuvre par la répétition des mêmes mots, la présence de coréférences et l'utilisation de mots sémantiquement reliés (Halliday and Hasan, 1976) et permet de suivre l'organisation des idées dans un texte. Dans le cadre de l'analyse de discours, (Xingwei, 1998) étudie les relations existant entre la cohésion et la cohérence au sein de documents textuels tandis que (Klebanov et al., 2008) développe une analyse stylistique de discours politiques. D'autres études reposent sur le principe de la cohésion lexicale pour des tâches de désambiguïsation du sens des mots (Manabu and Takeo, 1994) ou de résumés automatiques (Barzilay and Elhadad, 1997; Boguraev and Neff, 2000). Certains travaux d'identification et de correction d'erreurs se basent également sur la cohésion lexicale pour identifier des éléments qui ne sont pas sémantiquement reliés à leur contexte. Ces méthodes peuvent être appliquées sur du texte, comme dans (Hirst and Budanitsky, 2005), ou sur des transcriptions automatiques (Inkpen and Desilets, 2005). Finalement, de nombreuses études de segmentation thématique se fondent sur cette notion en employant une analyse de la distribution des mots au sein du texte pour détecter les ruptures thématiques (Utiyama and Isahara, 2001; Hearst, 1997).

Afin de développer des méthodes de structuration automatique de documents télévisuels capables de traiter plusieurs genres d'émissions, les techniques de structuration linéaire et hiérarchique développées dans cette thèse reposent sur la notion de cohésion lexicale. Contrairement aux conclusions présentées dans (Ostendorf et al., 2008), ce critère est malheureusement sensible aux particularités de nos données, présentées dans le chapitre 2. Nous avons, en effet, constaté sur nos émissions télévisuelles une dégradation importante des résultats obtenus par un algorithme de segmentation thématique, fondé sur la cohésion lexicale, sur les transcriptions automatiques par rapport aux transcriptions manuelles (*cf.* chapitre 5 pour plus de détail sur ces résultats). Cette dégradation des résultats s'explique par le fait que les transcriptions comportent des erreurs qui vont pénaliser le calcul de la fréquence des mots ; un mot n'est, en effet, pas détecté comme apparaissant plusieurs fois dans une transcription si ses multiples occurrences ne sont pas transcrites de la même façon. De plus, les documents audio-

visuels étudiés comportent peu de répétitions de vocabulaire du fait de l'utilisation massive de synonymes, dégradant ainsi les performances de la cohésion lexicale.

Pour pallier les difficultés liées aux spécificités des transcriptions automatiques de vidéos professionnelles, une mesure adaptée de ce critère de cohésion lexicale apparaît donc nécessaire. Dans ce cadre, certains travaux ont proposés d'ajouter à la seule notion de cohésion lexicale des indices propres aux documents oraux. Dans (Amaral and Trancoso, 2003) par exemple, les auteurs détectent la présence du présentateur dans les journaux télévisés pour améliorer la qualité de la segmentation thématique de journaux télévisés. Cependant de tels indices sont généralement très dépendants du type de documents traités. Dans cette thèse, nous proposons d'intégrer dans le calcul de la cohésion lexicale des informations provenant de sources linguistiques et acoustiques, indépendantes du domaine ou du genre des émissions télévisées étudiées. L'extraction de ces indices, employés dans le but de rendre le critère de cohésion lexicale plus robuste aux spécificités des transcriptions automatiques, est présentée dans la section 5.1.

Si le critère de cohésion lexicale, calculé sur les transcriptions automatiques, rend possible le développement de méthodes de structuration génériques, il ne permet pas de tirer profit de la multimodalité de nos données. Or, si les transcriptions obtenues à partir de la parole prononcée dans l'émission nous permettent d'accéder à son contenu sémantique, la façon dont est prononcée la parole au sein de l'émission est également un vecteur d'information important. Nous proposons donc, dans notre travail de thèse, d'intégrer des informations prosodiques lors du calcul de la cohésion lexicale. Ces informations acoustiques sont extraites d'une manière non supervisée, décrite dans la section 3.2, et sont donc totalement indépendantes du type des émissions traitées.

3.1 Gestion des spécificités des transcriptions automatiques de programmes TV

Pour pallier les erreurs de transcriptions et la faible répétition de vocabulaire au sein de nos données, nous intégrons dans nos méthodes de structuration des mesures de confiance, produites par le système de reconnaissance automatique de la parole, et des relations sémantiques. Nous décrivons dans cette section les méthodes d'extraction de ces deux types d'indices.

3.1.1 Mesures de confiance

Les mesures de confiance associées à chacun des mots de la transcription correspondent à la probabilité pour un mot d'être correctement transcrit. Ces mesures, générées *a posteriori* par le système de reconnaissance automatique à partir d'indices linguistiques, acoustiques et morphosyntaxiques, sont des informations qui peuvent s'avérer utiles pour pallier les erreurs de transcription présentes dans nos données. Elles sont couramment utilisées afin de détecter les erreurs au sein des transcriptions (Skantze and Edlund, 2004) ou les mots hors-vocabulaire (Sun et al., 2003; Lecouteux et al., 2009). Elles sont également retenues dans des tâches de plus hauts niveaux telles que la détection d'entités nommées (Miller et al., 1999; Favre et al., 2005).

Dans le cadre de cette thèse, les mesures de confiance vont être employées pour pénaliser les mots erronés en diminuant leur poids lors du calcul de la cohésion lexicale. Cependant, comme nous l'avons spécifié dans le chapitre 2, ces mesures de confiance ne sont pas toujours

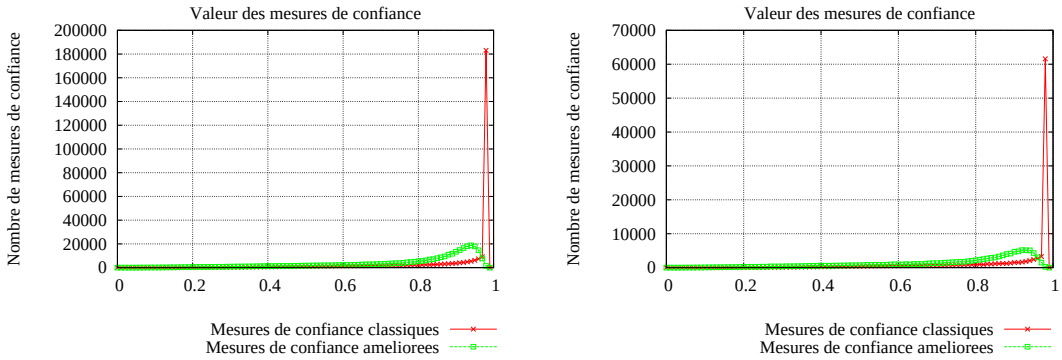


FIG. 3.1 – Distribution des mesures de confiance dans le corpus de journaux télévisés (a) et dans le corpus d'émissions *Sept à Huit* (b).

fiables et doivent être considérées avec prudence. En étudiant les valeurs des mesures de confiance associées aux transcriptions de nos corpora de journaux télévisés ou d'émissions de reportages *Sept à Huit*, nous constatons que leur distribution n'est pas équilibrée. Les valeurs de la majorité des mesures de confiance sont, en effet, comprises entre 0,98 et 1, ce qui ne traduit pas la réalité de la qualité des transcriptions – pour rappel la valeur du taux d'erreur mot sur les journaux télévisés est de VALEUR1 et de VALEUR2 sur les émissions *Sept à Huit*.

Dans (Fayolle et al., 2010), les auteurs proposent de générer des mesures de confiance plus fiables en se basant sur une combinaison de caractéristiques contextuelles des mots transcrits. Pour cela, ils prennent en compte des informations autres que les seuls scores acoustiques et linguistiques habituellement utilisés pour calculer ces mesures de confiance. Pour réévaluer la mesure de confiance d'un mot w, les auteurs étudient son contexte, les erreurs de transcription impactant non seulement w lui-même mais également les mots qui le précèdent ou qui le suivent. Ils font appel à des techniques d'apprentissage artificiel prenant en compte les caractéristiques phonétiques (durée et nombre de phonèmes) morpho-syntaxiques (l'étiquette catégorielle) ou linguistiques du contexte de w pour définir la nouvelle valeur de sa mesure de confiance. Les nouvelles mesures ainsi générées présentent une distribution plus équilibrée, les valeurs étant plus uniformément réparties entre 0,8 et 1, comme nous pouvons le constater sur la Figure 3.1.

Dans le cadre de notre travail de thèse, nous utilisons principalement les mesures de confiance générées directement par le système de reconnaissance automatique de la parole. Cependant, nous évaluons également l'impact de la qualité des mesures de confiance employées sur l'adaptation du critère de cohésion lexicale aux spécificités de nos données.

3.1.2 Relations sémantiques

L'utilisation de relations sémantiques pour adapter nos méthodes de structuration basées sur la cohésion lexicale sert deux objectifs. Premièrement, le critère de la cohésion lexicale se fondant uniquement sur la répétition des mots au sein d'un texte, sans prendre en considération le fait que deux mots différents peuvent être sémantiquement proches, nous souhaitons employer des relations sémantiques pour ajouter cette information dans nos méthodes de structuration, afin de gérer la faible répétition de vocabulaire de nos données (*cf.* chapitre 2). De plus, nous pensons que les relations sémantiques peuvent limiter l'impact des erreurs de

transcription. En effet, contrairement aux mots correctement transcrits, les mots mal reconnus sont peu susceptibles d'être sémantiquement liés à d'autres mots du document (Inkpen and Desilets, 2005). Par conséquent, l'utilisation des relations sémantiques doit diminuer l'impact des mots mal transcrits dans le calcul de la cohésion lexicale.

Il existe plusieurs types de relations sémantiques : les relations syntagmatiques et les relations paradigmatiques. Les relations syntagmatiques correspondent à des relations de successivité et de contiguïté que les mots entretiennent au sein d'une phrase (exemple : « conduire » et « voiture »). Le second type de relation réunit deux mots présentant une composante commune importante du point de vue du sens, comme « voiture » et « automobile ». Ces relations paradigmatiques regroupent des relations sémantiques dites hiérarchisantes – hyperonymie, méronymie, *etc.* – et des relations non hiérarchisantes comme la synonymie ou l'antonymie.

Selon Grefenstette, les méthodes automatiques d'extraction de relations sémantiques se décomposent en trois groupes (Grefenstette, 1994). Les techniques fondées sur les affinités dites de premier ordre, appelées par la suite « techniques de premier ordre », examinent le contexte local d'un mot pour déterminer les mots qui lui sont cooccurrents. En se basant sur l'hypothèse que les mots qui apparaissent souvent ensemble ont un lien de sens, ces méthodes permettent d'extraire des relations syntagmatiques. Les techniques qui permettent d'extraire des affinités de deuxième ordre définissent, quant à elles, le contexte de chaque mot et comparent ces contextes pour découvrir les mots similaires afin de retrouver des relations paradigmatiques. En effet, les mots synonymes, ou qui ont un sens proche, n'apparaissent généralement pas ensemble mais sont plutôt interchangeables et présentent donc des contextes similaires. Les méthodes consistant à découvrir des relations de troisième ordre tentent de caractériser le sens des mots en les regroupant le long d'axes sémantiques. Les trois groupes de techniques d'extraction de relations sémantiques sont présentés respectivement dans les trois premières parties de cette section. De plus, certains auteurs cherchent à aller plus loin que le troisième ordre défini par Grefenstette en étiquetant les relations existant entre les mots. Les méthodes développées pour caractériser le type des relations extraites sont décrites dans une quatrième partie. Finalement, dans une dernière section, nous présentons les relations sémantiques retenues pour nos travaux ainsi que leur méthodes d'extraction et de sélection.

3.1.2.1 Techniques de premier ordre

Les travaux ayant pour objectif de regrouper les mots d'un corpus de textes qui partagent un lien syntagmatique se basent généralement sur la détection de collocations. Ces collocations, c'est-à-dire ces groupes de mots apparaissant ensemble plus souvent que le hasard, sont repérées grâce à des calculs de fréquence d'apparition des mots dans le corpus, ensemble et séparément. Afin de mettre en relief cette notion d'apparition conjointe des composants dans le texte, chaque couple de mots d'un texte est associé à un score mesurant la force du lien unissant les deux mots, un score élevé traduisant un lien sémantique important. Si la méthode est simple, les scores utilisés peuvent être très variés. Parmi ceux utilisés dans ce cadre, un des plus connus est celui de l'information mutuelle proposé dans (Church and Hanks, 1989). Cette mesure compare la probabilité d'observer deux mots ensemble avec la probabilité de les observer séparément, et est calculée de la façon suivante :

$$I(w_1, w_2) = log_2 \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \ , \tag{3.1}$$

avec $P(w_1, w_2)$ la probabilité que les mots w_1 et w_2 soient au voisinage l'un de l'autre, c'est-à-dire dans une fenêtre de n mots. Cette probabilité est basée sur le nombre de fois où les deux

mots apparaissent dans un même contexte. $P(w_1)$ et $P(w_2)$ correspondent respectivement aux probabilités d'apparition des mots $w1$ et $w2$.

Certains travaux proposent d'appliquer et de modifier l'information mutuelle de différentes façons. Dans (Daille, 1994), l'auteur propose de mettre cette mesure au cube afin de donner plus de poids aux événements fréquents, la formule de base donnant, selon elle, trop de poids aux événements rares. Une indication sur l'ordre des mots peut également être intégrée à cette mesure de l'information mutuelle. Dans (Church and Hanks, 1990), les auteurs différencient la fréquence d'apparition du couple de mots $w_1 - w_2$ de celle du couple $w_2 - w_1$ et définissent ainsi ce qu'ils appellent l'*association ratio*. Dans (Brown et al., 1990), les auteurs utilisent l'information mutuelle pour créer ce qu'ils appellent les *sticky pairs* qui sont des paires de mots adjacents très fortement reliés, c'est-à-dire des couples de mots associés à une valeur d'information mutuelle très élevée. Dans ce travail, l'ordre des mots est également pris en compte ; ainsi « anciens combattants » est une *sticky pair* alors que « combattants anciens » n'en est pas une.

Le problème principal de ces méthodes est qu'elles ne fournissent que des collocations de deux mots. Dans (Brown et al., 1990), les auteurs proposent de construire une partition du vocabulaire du corpus en n classes de mots (n défini *a priori*) qui maximise la valeur moyenne d'un score similaire à l'information mutuelle. Ils obtiennent ainsi des classes qui regroupent des mots qui ont en commun une même racine (exemple : *performance*, *performed*, *perform*, *etc.*) et d'autres qui sont sémantiquement reliés sans avoir de racine commune (exemple : *attorney*, *counsel*, *trial*, *court* et *judge*). (Smadja, 1993) propose, quant à lui, d'étendre des collocations de deux mots à des collocations de n mots. Pour cela, il récupère toutes les phrases dans lesquelles apparaît une collocation de deux mots et l'étend si les mots qui l'entourent sont toujours les mêmes. Grâce à sa méthode, il extrait des collocations ainsi que des syntagmes rigides et des patrons qui peuvent être complétés – par exemple : *The NYSE's composite index of all its listed common stocks rose *NUMBER* to *NUMBER**, où *NUMBER* peut être remplacé par n'importe quel nombre. Dans (Frath et al., 1995), les auteurs se basent sur la fréquence de segments répétés, qui sont des suites de mots qui apparaissent souvent dans le corpus. Ces séquences sont simplifiées puis transformées en couples composés du premier et du dernier mot du segment répété. Le système Mantex ainsi mis en place regroupe les couples qui sont morphologiquement proches (ainsi le couple {appu*, touche*} représente les segments « appuyer longuement sur la touche » et « maintenez la touche appuyée »). Le système recherche ensuite les cooccurrences des couples dans des fenêtres de 20 mots. Les auteurs mettent ainsi en relation, non plus des mots, mais des couples de mots qui représentent des relations conceptuelles qui sont susceptibles de s'exprimer sous des formes différentes dans le texte.

3.1.2.2 Techniques de deuxième ordre

Les techniques de deuxième ordre, telles que définies par Grefenstette (Grefenstette, 1994), se basent sur l'observation suivante : deux mots sémantiquement proches ne sont pas utilisés l'un au voisinage de l'autre mais plutôt dans des voisinages similaires. Ces méthodes consistent donc à définir, d'une part, les propriétés contextuelles de chaque mot du corpus et, d'autre part, à mettre en relation, deux à deux, des mots partageant les mêmes propriétés contextuelles en calculant une mesure de similarité entre ces propriétés. Les propriétés considérées varient selon les travaux. Ce peut être les mots cooccurrant au sein d'une même fenêtre comme dans (Pichon and Sébillot, 1999) ou (Grefenstette, 1992). Dans ce cas, chaque mot plein et

lemmatisé d'un corpus est associé à un vecteur composé des mots apparaissant fréquemment dans son entourage, pondérés par leur fréquence d'apparition. La mise en relation de deux mots se fait en calculant la similarité entre leur vecteur représentatif. Cette similarité peut être calculée avec un indice de Jaccard, comme dans (Grefenstette, 1992), ou une mesure angulaire. Les propriétés contextuelles prises en compte peuvent également être des relations de dépendances au sein de syntagmes. Dans (Bourigault, 2002), Bourigault met en place le système UPERY qui construit des ontologies grâce à une analyse distibutionnelle. Dans un premier temps, cette analyse distributionnelle rapproche deux à deux les termes qui partagent des contextes similaires définis par l'analyseur syntaxique SYNTEX. La proximité entre deux termes est ensuite calculée à partir du nombre de contextes partagés par les deux mots, la productivité des contextes ainsi que le nombre de contextes propres à chacun des termes.

3.1.2.3 Techniques de troisième ordre

Les méthodes de troisième ordre ont pour objectif de comparer les mots similaires et de les regrouper le long d'axes sémantiques. Pour ce faire, Grefenstette (Grefenstette, 1994) établit les listes de mots similaires en comparant leurs contextes d'apparition avec un indice de Jaccard tel que mentionné dans la section 3.1.2.2. Ces listes, ordonnées du mot le plus proche au mot le moins proche sémantiquement, sont ensuite utilisées pour définir les axes sémantiques caractéristiques des différents sens du mot. Dans le même ordre d'idée, les travaux (de Chalendar and Grau, 2000; Pichon and Sébillot, 1999) cherchent à différencier les possibles sens d'un mot afin de ne pas regrouper dans une même classe sémantique des mots polysémiques. Pour cela, ces deux études se basent sur l'hypothèse qu'au sein d'un même domaine sémantique les mots ne possèdent plus qu'un seul sens – le mot *souris* n'étant plus polysémique lorsque le contexte dans lequel il apparaît traite de l'informatique. Dans (de Chalendar and Grau, 2000), le système SVETLAN' regroupe les noms jouant un même rôle syntaxique pour un même verbe au sein de textes traitant d'un même domaine sémantique. Dans (Pichon and Sébillot, 1999), les auteurs amorcent également leur travail sur des textes thématiquement homogènes. Pour étudier les différences existant entre les sens d'un mot w, ils extraient, pour chacun des thèmes du corpus dans lequel il apparaît, une liste d'environ 20 mots présents au moins 2 fois dans le voisinage de w. Ces listes sont ensuite étudiées afin de faire ressortir les caractéristiques des différents sens de w. Dans (Rossignol and Sébillot, 2006) les auteurs poussent un peu plus loin cette étude en proposant une automatisation plus importante du processus. La technique mise en point dans ce travail repose sur l'étude de grands voisinage autour des occurrences des mots comparés pour réaliser un regroupement de paires de mots, distingués par des nuances similaires, « abstrait/concret » par exemple. Certaines mesures ont également été définies pour tenter de mettre en avant les différences existant entre deux mots sémantiquement proches. Dans (Church et al., 1991), les auteurs présentent la mesure *t-score* qui permet de différencier deux mots comme *powerful* et *strong* en fournissant les mots qui apparaissent significativement plus souvent après *strong* qu'après *powerful* et *vice-versa*.

3.1.2.4 Caractérisation des relations sémantiques

Pour finir, certains travaux poussent plus loin l'objectif défini par Grefenstette et proposent des méthodes pour étiqueter les relations sémantiques. En effet, les techniques présentées jusqu'à présent fournissent généralement des couples de mots, associés à un score de proximité sémantique, qui peuvent être reliés par des types de relations très différentes (synonymie, hyperonymie, antonymie, *etc.*) sans que cette différence ne soit explicitée. Dans (Hearst, 1992;

Morin, 1998), les auteurs présentent une méthode d'extraction de patrons lexico-syntaxiques caractéristiques d'un type de relation à partir d'une amorce constituée de couples de mots qui respectent cette relation. Cette méthode, appliquée sur l'anglais et le français fournit de bons résultats pour la relation d'hyperonymie. Cependant, lorsqu'elle est appliquée à un autre type de relations sémantiques – la méronymie – les résultats sont nettement moins bons, ce qui s'explique, selon Hearst, par le fait que la relation de méronymie s'exprime de façon plus variée et moins figée que celle d'hyperonymie. Certains travaux proposent également d'étiqueter les relations sémantiques en amorçant leur analyse sur des classes obtenues lors d'une phase préalable. Dans (Daille, 2003) par exemple, l'auteur met en place une méthode regroupant dans une même classe et caractérisant les différentes variantes d'un même terme complexe. Les termes complexes et les variations sont extraits d'un corpus grâce à un système de règles. Le lien unissant un terme complexe à l'une de ces variantes est ensuite caractérisé grâce à des indices syntaxiques (l'ajout ou la présence d'un adjectif relationnel marque une relation d'hyperonymie alors que l'ajout d'un adverbe négatif marque une relation d'antonymie) ou à des indices morphologiques. Finalement, certaines études extraient la relation de synonymie en utilisant un alignement entre corpora multilingues (Dyvik, 2002; Muller and Langlais, 2010). L'idée sous-jacente à ces travaux est que si deux mots sont reliés par une relation de synonymie, alors il partagent probablement largement leur traduction. En comparant l'extraction automatique de synonymes par une méthode distributionnelle classique et par une méthode basée traduction, (Muller and Langlais, 2010) montre que la technique d'alignement fournit de meilleurs résultats que la méthode distributionnelle.

3.1.2.5 Relations utilisées dans cette thèse

L'objectif de cette thèse étant d'étudier l'influence de l'intégration de relations sémantiques et non d'optimiser leur extraction, nous avons choisi d'appliquer des méthodes standards afin d'acquérir deux types de relations : des relations syntagmatiques et des relations paradigmatiques. Pour calculer les relations syntagmatiques, nous avons retenu deux indices de force d'association couramment utilisés : l'information mutuelle IM et l'information mutuelle au cube IM^3 (Daille, 1994). L'extraction des relations paradigmatiques est réalisée en associant à chaque couple de mots le cosinus de l'angle entre les vecteurs de voisinage des occurrences des deux mots. Le vecteur de voisinage d'un mot w est composé des mots pleins lemmatisés apparaissant dans une fenêtre de 20 mots centrée sur w. Nous obtenons ainsi une liste de synonymes, d'hyperonymes, *etc.*, non différenciés.

Pour l'ensemble des méthodes d'acquisition, les relations ont été extraites sur un corpus composé d'articles *du Monde*, *de l'Humanité* et des transcriptions de référence des campagnes ESTER 1 et ESTER 2 correspondant respectivement à 100 et 150 heures de journaux radiophoniques. Dans ce corpus, lemmatisé et normalisé, seuls les noms, adjectifs, et verbes autres que « être », « avoir » et « falloir » ont été conservés. Les scores d'association ont finalement été normalisés afin d'obtenir des valeurs comprises entre 0 et 1.

La question qui se pose alors est de retenir, pour notre tâche de segmentation, les relations sémantiques les plus pertinentes parmi tous les liens extraits. Nous explorons deux méthodes de sélection :

- $Total_{\rho_1}$ qui consiste à conserver, pour chaque mot du corpus d'apprentissage, les ρ_1 relations ayant les scores les plus élevés ; ces relations sont appelées « premières relations » dans la suite de ce manuscrit. Dans nos tests, la valeur de ρ_1 peut être égale à 5 000, 10 000, 20 000, 50 000 et 90 000 ;

TAB. 3.1 – Relations aux scores d'association les plus élevés pour le mot « cigarette »

	IM	IM^3	paradigmatique
$Total_{90000}$		cigarette fumer cigarette paquet	cigarette cigare cigarette gitane cigarette gauloise
$ParMot_{10}$	cigarette chevignon cigarette liggett cigarette altadi cigarette détaxer	cigarette fumer cigarette paquet cigarette allumer cigarette contrebande	cigarette cigare cigarette gitane cigarette gauloise cigarette clope

– $ParMot_{\rho_2}$ dans laquelle un nombre fixe ρ_2 des meilleures relations associées à chaque mot du corpus sont conservées, ρ_2 prenant les valeurs 2, 3 et 10.

De plus, nous avons remarqué que certains mots du corpus d'apprentissage, comme « aller », « an », *etc.*, étaient sémantiquement liés avec un nombre important d'autres mots. Afin d'éviter de créer des liens sémantiques entre de trop nombreux couples de mots, nous avons défini une technique de filtrage, $Seuil_\gamma$, pouvant être associée aux deux méthodes de sélection. Elle consiste à ignorer les relations sémantiques des mots qui entretiennent un nombre de relations supérieur à un certain seuil. La valeur du seuil est égale au nombre moyen de relations associées aux mots d'un document multiplié par un paramètre γ prenant des valeurs comprises entre 0 et 1.

Le tableau 3.1 illustre les 5 relations aux scores d'association les plus élevés rattachées au mot « cigarette » en sélectionnant les 90 000 premières relations, $Total_{90000}$, et 10 relations par mot, $ParMot_{10}$. Nous constatons que la qualité de ces relations change selon la méthode utilisée pour leur extraction. Ainsi, les relations obtenues grâce au score *IM* correspondent généralement à des événements rares, à tel point d'ailleurs qu'aucune des relations existant avec le mot « cigarette » n'apparaît pas dans les 90 000 premières relations, alors que les relations IM^3 et paradigmatiques semblent plus pertinentes. De ce fait, dans la suite de notre travail, les relations syntagmatiques employées sont celles obtenues grâce à la mesure IM^3.

3.2 Utilisation de la prosodie

Si notre tâche de structuration de documents multimédias est principalement mise en œuvre à travers les transcriptions automatiques de la parole contenues dans les émissions, il est important de prendre en compte un maximum d'indices offerts par la multimodalité de nos données, tout en conservant à l'esprit notre souci de généricité. Dans cette optique, nous avons choisi d'utiliser la prosodie, d'une part, parce qu'elle peut être extraite de façon non supervisée et, d'autre part, parce que cette information est un élément crucial de la communication parlée.

La relation qui existe entre les mots prononcés et leur prosodie est souvent comprise comme étant la différence entre *ce que l'on dit* et *comment on le dit*. Là où les mots décrivent le contenu lexical de la parole prononcée, la prosodie décrit la manière dont ces mots sont prononcés. La prosodie peut refléter diverses caractéristiques de l'énonciation : l'état émotionnel du locuteur ; si un énoncé est une déclaration, une question, ou un ordre ; si le locuteur est ironique ou sarcastique ; s'il met l'accent sur un énoncé. De plus, le fait qu'un énoncé soit accentué, ou mis en évidence dans le discours de façon intentionnelle, peut indiquer que l'information portée par cet énoncé est une information nouvelle ou importante du point de vue du locuteur

(Hirschberg, 2002). En effet, dans (Crestani, 2001), l'auteur montre qu'il existe, dans la langue anglaise, un lien direct entre l'accentuation acoustique d'un terme et sa valeur informative.

Nous proposons dans cette thèse d'étudier l'impact de la prosodie sur des techniques de segmentation thématique et de mise en relation de segments thématiques de documents en langue française.

Afin d'associer des informations acoustiques à nos données télévisuelles, deux caractéristiques sont extraites de chaque mot plein d'un document : l'énergie (*intensité*), présentant le niveau sonore perçu, et la fréquence fondamentale (*pitch*), traduisant la proéminence d'une syllabe ou d'un mot prononcé par un locuteur. Le choix de ces deux critères repose sur le fait qu'il existe une corrélation entre l'augmentation de leur valeur pour un mot donné w et le statut informatif de ce mot.

Pour chaque document de notre corpus, les valeurs d'intensité et de pitch sont extraites pour chaque fenêtre de 0,01 seconde du fichier audio par le biais du logiciel Praat (Boersma and Weenink, 2002). Les valeurs sont ensuite normalisées pour chaque locuteur grâce à un z-score. La segmentation en locuteurs de l'émission est générée par le système de reconnaissance automatique de la parole IRENE décrit dans le chapitre 2. Cette segmentation en locuteurs étant imparfaite[1], cette normalisation reste approximative.

Les transcriptions automatiques sont ensuite alignées avec les informations acoustiques grâce aux limites temporelles associées à chacun des mots par le système de reconnaissance automatique de la parole. Un mot étant généralement associé à plusieurs valeurs d'intensité et de pitch – la durée de prononciation d'un mot est globalement supérieure à 0,01 seconde –, quatre stratégies ont été utilisées pour calculer le score d'un mot. La première stratégie, MAX, consiste à associer au mot la valeur maximale observée pour chaque fenêtre de 0,01 seconde. Pour les techniques MOYENNE et ET, la moyenne et l'écart-type sont calculés, à partir des scores de chacune des fenêtres, puis associés à w. Finalement, la stratégie MIN consiste à conserver la valeur minimum d'intensité et de pitch.

Chaque mot w apparaissant dans les transcriptions est ainsi associé à deux scores $i(w) \in [0, 1]$ et $p(w) \in [0, 1]$ traduisant respectivement l'intensité et le pitch de ce mot. Une troisième valeur $b(w)$ combinant ces deux scores est également calculée. Afin de pénaliser fortement les mots associés à une faible valeur de d'intensité ou de pitch, la multiplication entre ces deux valeurs a été préférée au calcul de la moyenne.

3.3 Bilan du chapitre

Ce chapitre, ainsi que les précédents, nous ont permis de présenter les bases sur lesquelles repose notre travail de thèse : les méthodes de structuration mises en place, les spécificités de nos données et les indices utilisés pour pallier ces particularités. Avec le chapitre suivant s'ouvre une nouvelle partie de notre thèse consacrée à la segmentation thématique linéaire. Dans le chapitre 4, sont exposées les différentes techniques de segmentation thématique existantes ainsi que la méthode retenue dans notre travail. Cette méthode est ensuite adaptée aux transcriptions automatiques de vidéos professionnelles, grâce à l'utilisation des indices présentés dans ce chapitre. L'intégration de ces informations et leur impact sont décrits dans le chapitre 5. Les méthodes de structuration proprement dites sont, finalement, exposées dans la dernière partie du manuscrit, correspondant aux chapitres 6 et 7.

[1]Bien que les segments partageant une même classe sont associés à un même locuteur, l'ensemble des segments prononcés par un locuteur donné peut être réparti dans plusieurs classes.

Deuxième partie

Segmentation thématique

Chapitre 4

Détection de rupture et maximisation de la cohésion lexicale pour la segmentation thématique linéaire : état de l'art et positionnement

La segmentation thématique linéaire consiste à mettre en évidence la structure sémantique d'un document. Pour cela, les algorithmes développés cherchent à détecter de façon automatique les frontières délimitant les segments thématiquement cohérents. Si la segmentation thématique d'un texte constitue un objectif en tant que tel, elle est également utilisée comme point de départ de nombreux travaux de traitement automatique des langues, notamment en résumé automatique. Dans ce cadre, certains auteurs, (Angheluta et al., 2002; McDonald and Chen, 2002) ou (Boguraev and Neff, 2000) entre autres, réalisent une étape préalable de segmentation thématique, autorisant l'extraction des différents éléments abordés dans le document, dans le but de dégager les principaux faits à inclure dans le résumé. De même, certaines études de recherche d'information passent par une phase de segmentation thématique afin de comparer la requête fournie par l'utilisateur à des parties thématiquement cohérentes extraites des documents, et améliorer, ainsi, la mesure de similarité entre documents et requête (Moffat et al., 1994; Chan, 2000). De plus, les auteurs de (Hearst and Plaunt, 1993) ont montré qu'indexer des segments thématiquement homogènes permettait d'accroître les performances d'une tâche de recherche d'information comparativement à l'indexation classique. Finalement, la segmentation thématique est employée, comme nous l'avons vu dans le chapitre 1, pour la détection et le suivi d'événements au sein de documents télévisuels (Wu et al., 2006; Ide et al., 2005). De façon analogue à ces travaux, la phase de segmentation thématique linéaire développée dans cette thèse peut constituer la première étape des méthodes de structuration automatique de documents télévisuels.

De très nombreux travaux se sont penchés sur la tâche de segmentation thématique linéaire, de documents textuels ou audiovisuels, en employant des indices de natures variées : linguistique (Litman and Passonneau, 1995), acoustique (Hirshberg and Nakatani, 1998), *etc.* Cette thèse se caractérisant par une volonté de développer des techniques applicables à différents types d'émissions, nous avons souhaité utiliser une méthode de segmentation thématique fondée sur la cohésion lexicale, n'exploitant donc pas des indices trop spécifiques à certains programmes. La cohésion lexicale est à la base de nombreuses études, qui peuvent se diviser en

deux grandes familles. Les méthodes locales, qui consistent à détecter localement les ruptures de la cohésion lexicale apparaissant dans le document pour définir les frontières thématiques, et les méthodes globales, qui reposent sur la mesure de la cohésion lexicale et cherchent à maximiser la valeur de la cohésion lexicale des segmentations obtenues. Les deux types de méthodes de segmentation, habituellement employés indépendamment l'un de l'autre, étant à nos yeux complémentaires, l'objectif de ce chapitre est double. Tout d'abord, nous présentons dans la section 4.2 un état de l'art des différentes techniques de segmentation thématique, qu'elles soient locales ou globales. Puis, nous proposons une approche originale consistant à fusionner les deux types de méthode afin d'obtenir une segmentation thématique capable de fournir des segments à la fois très cohérents et très différents les uns des autres. Cette approche est décrite dans la section 4.3.

Cependant, avant de rentrer dans le détail des techniques de segmentation, nous souhaitons aborder la problématique de la définition du thème. En effet, bien que de nombreux travaux aient cherché à formaliser cette notion, elle reste extrêmement intuitive. Nous commençons donc ce chapitre par un panorama des différentes formalisations proposées dans la littérature, avant de nous intéresser, plus particulièrement, à la définition du thème dans le cadre de la segmentation de données audiovisuelles et de présenter celle adoptée dans cette thèse.

4.1 Thème : définition

La notion de thème, si elle est au centre de nombreux travaux d'informatique linguistique, est rarement définie précisément, et, lorsqu'elle l'est, c'est souvent par le biais d'une paraphrase telle que « ce dont on parle » ou « ce dont parle l'énoncé ». Une définition claire et non pas seulement intuitive semble pourtant nécessaire pour cerner correctement les problèmes à résoudre – segmentation thématique, caractérisation des thèmes présents dans un document – et évaluer les résultats obtenus par les méthodes développées. De nombreux linguistes se sont penchés sur une conceptualisation formelle de cette notion et tous ces travaux conduisent à une quantité importante de définitions. Dans une première partie, nous cherchons à donner un aperçu des différentes formalisations proposées en mettant en avant les principaux points d'accord existant entre tous ces travaux. Puis, nous abordons la définition du thème à travers le cadre d'analyse qu'est la segmentation thématique de données multimédias ; dans cette seconde partie, nous décrivons les définitions existantes, ainsi que celle que nous avons choisi d'utiliser dans le cadre de cette thèse.

4.1.1 Définition du thème dans la littérature

De nombreux linguistes ont cherché à caractériser le thème, proposant ainsi un nombre très important de définitions. Le thème est en effet associé à un agglomérat de notions et le simple fait d'opter pour un système de dénomination pose déjà un problème. On peut ainsi trouver dans la littérature les termes de : *topos*, *topic*, *thème*, *sujet*, *motif*, *etc.* Ces termes, s'ils recouvrent parfois des notions clairement différenciées par les auteurs, sont souvent utilisés de façon intercheangeable. Nous présentons ici une liste – non exhaustive – des différentes définitions qu'il est possible de trouver dans la littérature. Le thème est ainsi défini comme :

- une position syntaxique,
- le premier élément de l'énoncé,
- l'idée qu'une information est communiquée à propos de x (expressions interprétantes : *à propos de*, *ce dont on parle*),

- l'idée d'importance ou de saillance d'un objet ou d'un sujet dans la conscience d'un locuteur ou dans son discours,
- un élément assurant la cohérence de la suite des énoncés,
- l'objet central d'une description,
- une information connue ou donnée en regard du savoir partagé des interlocuteurs,
- *etc.*

Face à cette multitude de définitions, certains auteurs ont cherché à résumer et à regrouper les différentes caractérisations existantes. C'est le cas, par exemple, d'Anne Grobet qui, dans (Grobet, 2002), propose une cartographie des définitions du thème, qu'elle nomme par ailleurs indifféremment thème ou topique, qui lui permet de mettre au jour cinq grands types de définitions. Le thème peut ainsi être défini comme :

1. l'information donnée. Dans ce cas, le thème est assimilé à l'information présentée par le locuteur (ou auteur d'un texte) comme connue par l'interlocuteur (ou le lecteur) ;
2. ce dont parle l'énoncé. Le thème, ce dont on parle, est ici distingué du propos, ce qu'on en dit ;
3. le point de départ de l'énoncé ;
4. l'élément porteur du plus bas degré de dynamisme communicatif. Dans cette définition, qui se base non plus sur un couple oppositif mais sur un *continuum* informatif caractérisant les éléments de l'énoncé, le thème est vu comme l'élément le moins informatif ;
5. le thème discursif qui est, selon Grobet, souvent exclu des discussions terminologiques. Le thème constitue, dans ce cas, un principe organisateur qui résume et structure la représentation sémantique du discours qui peut être paraphrasée par « ce dont parle le discours ».

Cette tentative de cartographie des définitions du thème est révélatrice de la difficulté de traiter de la notion du thème. En effet, Grobet précise que la distinction proposée est grossière et note que les différentes définitions ne sont pas nécessairement exclusives. Ainsi, on constate que la distinction entre le thème vu comme ce dont parle l'énoncé et comme point de départ de l'énoncé est assez faible. De plus, si Grobet fait ici une différenciation entre thème (correspondant au thème d'un énoncé, généralement opposé au rhème) et thème discursif, c'est rarement le cas dans les définitions existantes. Enfin, si le fait d'avoir de nombreuses définitions n'est pas problématique en soi – une même notion pouvant être définie de différentes façons, ou sous différents points de vue (sémantique, paradigmatique, logique, *etc.*) –, la difficulté rencontrée dans la conceptualisation du thème vient du fait que toutes ces définitions ne sont pas clairement indépendantes. Cependant, s'il existe de nombreuses différences ou incohérences entre les formalisations établies, certaines définitions partagent tout de même de nombreux points communs.

Par exemple, dans leur travaux, Brown et Yule (Brown and Yule, 1983) soulignent la part de subjectivité inhérente à l'identification du thème d'un discours qui dépend des points de vue des interlocuteurs. Or, cette idée de subjectivité est reprise par de nombreux auteurs et est sans doute à la base de la difficulté de formalisation de la notion de thème. Le sens d'un texte n'est pas figé et la compréhension qu'en ont les lecteurs est évolutive. Le thème est ainsi fonction de subjectivité et dépend de la sensibilité et de la culture des lecteurs. En effet, bien que Litman et Passonneau aient noté dans (Passonneau and Litman, 1993) une concordance inter-annotateurs importante dans le placement des frontières de segments thématiques, entre 82 et 92 %, elles ont également mis en évidence de fortes irrégularités dans les taux de placement des frontières entre les différents utilisateurs. Ces taux de placement de

frontières, variant de 5,5% à 41,3%, ainsi que la variation importante de la taille des segments résultats, mettent en évidence la difficulté de définir une seul segmentation pour un même document. La différence de segmentation semble venir d'une différence de perception sur la notion de thème et notamment de la granularité de cette notion. Là où certains annotateurs considèrent qu'un seul thème est abordé dans un segment, d'autres y perçoivent des nuances et proposent une segmentation de granularité plus fine mettant en avant différents sous-thèmes plus précis.

Selon Marandin (Marandin, 1988), thématiser « c'est stabiliser un état du monde raconté et se satisfaire du monde partiel ». Il s'agit de déterminer comment une suite d'énoncés permet des lectures différentes et ce qui, dans les énoncés ou dans leur enchaînement, oriente vers telle lecture plutôt que vers telle autre. Pour lui, la compréhension d'un texte est vue comme un processus d'augmentation où le thème serait la toile de fond. Le thème vu comme base cohérente nécessaire à la compréhension est commun à un certain nombre de définitions. Ainsi, chez Rastier[1] (Rastier, 1995), le thème est une structure stable au sein de laquelle il n'y a pas de polysémie ; pour Grobet (Grobet, 2002), c'est une condition de pertinence pour l'interprétation des énoncés. Chez Fradin et Cadiot (Fradin and Cadiot, 1988), le thème est un point de cohérence entrevu ou supposé. Cette notion est également présente dans des définitions qui voient le thème comme une information connue ou toutes les définitions intuitives telles que « ce dont on parle ».

Finalement, de nombreux auteurs s'accordent à dire que les définitions proposées ne prennent tout leur sens que lorsqu'elles sont formulées dans un cadre d'analyse très précis. En effet, il est clair qu'il n'est pas possible de formuler une définition qui soit valable à tous les niveaux, pour tous les types de textes. Dans la section suivante, nous nous attachons donc à étudier les définitions proposées dans un cadre d'analyse précis proche du nôtre qui est la détection de thèmes dans les émissions télévisuelles.

4.1.2 Le thème dans le cadre de données audiovisuelles

Le projet *Topic Detection and Tracking*, organisé par le NIST depuis 1997, a pour but de retrouver des segments corrélés thématiquement au sein de flux multimédias traitant de l'actualité : radio d'information en continu, journaux télévisés, *etc.* Ainsi, les participants au projet cherchent à associer les segments de ces flux, issus de plusieurs sources, qui traitent d'un même sujet d'actualité. Afin d'évaluer les méthodes mises en place, les organisateurs ont coordonné la création d'une vérité terrain par le biais d'un guide fourni à tous les annotateurs. Ce guide propose une définition des termes *topic* et *event*, définitions qui sont reprises dans de nombreux articles traitant de la détection ou du suivi de sujet dans l'actualité, par exemple (Yang et al., 2000) ou (Fukumoto and Suzuki, 2000).

Un *event* est défini comme étant un élément spécifique localisé dans l'espace et dans le temps, ainsi que toutes ses préconditions et ses inévitables conséquences. Par exemple, en Février 1998, un avion de l'armée américaine volant à basse altitude a percuté un câble supportant un téléphérique à la station de ski Calavese en Italie. Le téléphérique est ensuite tombé au sol, tuant 20 personnes et en blessant de nombreuses autres. La chute du téléphérique ainsi que les morts et les blessés qu'elle a engendrée sont les conséquences inévitables de la collision entre l'avion et le câble ; elles sont donc considérées comme faisant toutes partie du même *event*. Un *topic*, quant à lui, est un *event* accompagné de tous les *events* qui y sont

[1] Cf. Annexe A pour une description plus détaillée de la définition du thème proposée par Rastier (Rastier, 1995) et Marandin (Marandin, 1988).

directement reliés. Pour passer de l'*event* au *topic*, il faut donc être capable de définir les *events* directement reliés à un *seminal event*, c'est-à-dire à l'événement principal – la collision entre l'avion de l'armée américaine et le câble soutenant le téléphérique dans notre exemple. Les secours mis en place ou les funérailles des victimes sont ainsi des *events* directement reliés au premier et ils appartiennent donc au même *topic*. Pour simplifier la tâche des annotateurs lorsqu'ils cherchent à trouver les *events* liés à un *seminal event*, les organisateurs ont mis en place 13 règles d'interprétation. Ces règles stipulent, pour chaque *seminal event*, les types d'*events* qui peuvent être considérés comme reliés. La différence majeure existant entre un *topic* et un *event* est que l'*event* est relativement court et évolue dans le temps contrairement au *topic* qui est plus stable et long. Cette notion de stabilité du *topic* peut être rapprochée de celle retrouvée communément dans les différentes définitions du thème décrites dans la section précédente.

Dans le cadre de notre travail de thèse, notre tâche de segmentation consiste à découper les journaux télévisés ou émissions de reportages qui constituent notre corpus en unités qui se rapprochent des *events* définis dans le cadre de TDT. En effet, nous allons, dans le cadre de la segmentation thématique linéaire, segmenter nos émissions en reportages. Ces reportages sont composés des trois éléments suivants :

- le plateau de lancement durant lequel le journaliste présente le contenu de l'*event*,
- le reportage filmé,
- le retour plateau contenant la conclusion du journaliste.

Ces deux derniers éléments sont facultatifs, le lancement plateau pouvant constituer la seule référence à l'*event* dans l'émission, notamment dans les journaux télévisés. Si dans le cadre des journaux télévisés le contenu des reportages correspond bien à la définition d'un *event* proposée par TDT, il diffère légèrement dans le corpus d'émissions de reportages, les sujets abordés dans ce type d'émissions étant plus vaste qu'un simple fait d'actualité sans pour autant se rapprocher forcément de la définition de *topic*.

De fait, afin de proposer un objectif commun à notre tâche de segmentation thématique linéaire d'émissions télévisuelles, un segment thématiquement cohérent correspond, dans notre thèse, à un reportage (éventuellement associé à ces plateaux de lancement et de fin) qui constitue une unité sémantique dans la structure éditorialiste de l'émission.

4.2 Segmentation thématique

Comme nous l'avons signalé précédemment, la segmentation thématique de documents, qu'ils soient textuels ou audiovisuels, a fait l'objet d'un nombre très important de travaux. Parmi les techniques mises en place, on distingue les méthodes supervisées des non supervisées, celles qui prennent en compte des marqueurs linguistiques (Beeferman et al., 1997; Litman and Passonneau, 1995), des indices vidéos (Amaral and Trancoso, 2003), ou une combinaison de ces différentes informations.

Cependant beaucoup de ces travaux sont inadaptés à notre cadre d'étude, les indices utilisés les rendant très spécifiques à un type de corpus ou de documents. Afin de développer une méthode de segmentation thématique capable de traiter des émissions télévisuelles tout-venant, nous choisissons de fonder notre technique de segmentation thématique sur le critère de la cohésion lexicale. Cette cohésion lexicale peut être mesurée grâce à des méthodes statistiques ce qui la rend tout à fait adaptée à nos données, d'une part parce qu'elle ne nécessite pas de phase d'apprentissage et, d'autre part, parce qu'elle est indépendante du type de documents

considérés.

4.2.1 Segmentation thématique fondée sur la cohésion lexicale

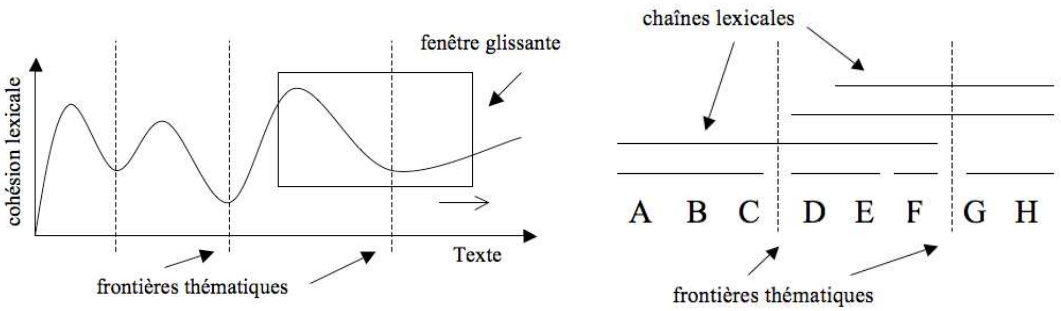
De fait, ce critère est très populaire et constitue la base de nombreuses techniques de segmentation thématique. Ces méthodes peuvent se diviser en deux grandes familles. Premièrement, les méthodes, que nous qualifions de *locales*, se basent sur la détection de la *rupture de la cohésion lexicale* en étudiant localement la valeur de ce critère. Deuxièmement, les techniques *globales* sont fondées sur la *mesure de la cohésion lexicale* et cherchent à obtenir les segmentations maximisant cette cohésion.

4.2.1.1 Méthodes locales fondées sur la détection de rupture de la cohésion lexicale

L'algorithme TextTiling, présenté dans (Hearst, 1997), est un algorithme fondateur de la segmentation thématique qui repose sur l'analyse de la distribution des mots au sein du texte ; un changement de vocabulaire important étant considéré comme une marque de changement thématique. Cet algorithme propose une technique de segmentation thématique, fondée sur un principe de fenêtre glissante, qui sera reprise dans de nombreux travaux. Dans ce travail, le texte à segmenter est parcouru par une fenêtre de longueur paramétrable centrée en un point du texte, correspondant à une frontière thématique potentielle (frontière entre deux phrases, deux paragraphes, *etc.*). Les parties droite et gauche de la fenêtre sont représentées grâce à des vecteurs pondérés caractéristiques de leur contenu. Une mesure de similarité est calculée entre ces deux vecteurs afin d'obtenir une valeur de la cohésion lexicale, traduisant la ressemblance lexicale entre les deux parties de la fenêtre. La fenêtre est ensuite décalée dans le texte pour calculer une valeur de la cohésion lexicale au niveau d'une autre frontière thématique potentielle. Une courbe tracée à partir de ces valeurs permet d'extraire les frontières thématiques en identifiant les vallées les plus profondes (*cf.* Figure 4.1), c'est-à-dire les positions du texte où la cohésion lexicale est la plus faible, correspondant aux changements de vocabulaire les plus marqués[2]. Dans (Hearst, 1997), l'auteur calcule une valeur de la cohésion lexicale entre des blocs de pseudo-phrases qu'il préfère à des paragraphes. Hearst considère en effet que la différence de longueurs existant entre les paragraphes va pénaliser le calcul de la mesure de similarité. Les vecteurs représentatifs des deux parties de la fenêtre sont composés des mots pleins lemmatisés, pondérés par leur fréquence, et la mesure utilisée est la mesure angulaire cosinus.

Cet algorithme simple à mettre en œuvre a inspiré de nombreux travaux qui ont cherché à améliorer les résultats obtenus en modifiant la représentation vectorielle ainsi que la mesure de similarité calculée. Dans (Hernandez and Grau, 2002) par exemple, les auteurs utilisent une pondération *tf-idf* et un produit scalaire pour calculer les valeurs de la cohésion lexicale. De plus, ils évaluent la cohésion lexicale entre deux paragraphes, plutôt que sur des blocs de pseudo-phrases comme proposé par (Hearst, 1997) car, selon eux, les auteurs d'un texte ont tendance à exposer un point de vue par paragraphe, qui constitue donc une entité thématiquement homogène. Dans (Ferret et al., 1998), l'estimation de la similarité se base sur une mesure de *Dice*, calculée entre deux paragraphes (représentés par des vecteur de mots pleins lemmatisés pondérés par le poids *tf-idf*). Les auteurs proposent également de prendre en compte

[2]Typiquement une frontière thématique est choisie lorsque la valeur de la cohésion lexicale est égale à la moyenne des valeurs de la cohésion lexicale moins la moitié de l'écart-type.



(a) (b)

FIG. 4.1 – Principe de la segmentation thématique locale à base de fenêtre glissante (a) et de chaînes lexicales (b).

des relations sémantiques pour pallier une faible répétition de vocabulaire liée à la présence de synonymes dans leurs données. Pour cela, ils intègrent des relations sémantiques, obtenues automatiquement à partir d'un corpus composé d'articles du *Monde*, dans le calcul des pondérations associées aux mots des vecteurs caractéristiques. Pour chaque mot w du vecteur, le poids des mots du paragraphe, reliés sémantiquement à w, est augmenté proportionnellement au nombre d'occurrence de w et à la force du lien unissant les deux mots dans le réseau de collocations. Dans (Ferret, 2002), l'auteur se base également sur l'utilisation d'un réseau de relations sémantiques pour aider la segmentation linéaire. Pour cela, il déplace un fenêtre glissante sur un texte et lui associe un contexte thématique. Les segments du texte sont également associés à un contexte thématique[3]. La segmentation se fait alors de la façon suivante : si la similarité (calculée grâce à une mesure cosinus) entre le contexte de la fenêtre glissante et celui du segment actif est faible, alors l'algorithme détecte un changement thématique et le segment actif est clos, sinon le segment actif est étendu afin d'englober la position courante de la fenêtre glissante. Dans (Claveau and Lefèvre, 2011a), les auteurs proposent une technique de comparaison indirecte permettant de mettre en évidence une similarité sémantique entre deux paragraphes d'un texte, même si ceux-ci ne partagent pas de vocabulaire. Cette méthode consiste à utiliser des vecteurs composés de m scores traduisant la proximité sémantique entre une partie du texte caractérisée par le vecteur et m documents-pivots. L'idée sous-jacente à ce travail est que si deux parties du texte à segmenter sont sémantiquement proches des mêmes documents, elles ont un lien sémantique important. Finalement, les auteurs de (Labadié and Chauché, 2007) choisissent une représentation vectorielle des phrases construite en projetant chaque mot dans un espace de concepts de dimension finie permettant de leur associer un vecteur de concepts. Une représentation syntaxique de la phrase, sous la forme d'un graphe, est ensuite utilisée pour inférer le vecteur de la phrase à partir des vecteurs des mots qui la constitue. La similarité entre les vecteurs est finalement calculée grâce à l'*arc cosinus*. Cette fonction, fortement non linéaire pour des valeurs d'angles faibles, se comporte de manière quasi linéaire pour les valeurs d'angles élevées et offre ainsi une analyse plus fine lorsque deux phrases sont sémantiquement proches.

Afin de prendre en compte les relations existant entre les différentes occurrences des mots

[3]Le contexte thématique de la fenêtre (*resp.* des segments) est constitué à la fois des mots de la fenêtre (*resp.* des segments) et des mots d'un réseau de collocations jugés les plus fortement reliés aux mots de la fenêtre (*resp.* des segments).

apparaissant dans le vocabulaire du texte à segmenter, plusieurs travaux ont proposé d'utiliser les chaînes lexicales pour représenter les deux parties de la fenêtre glissante. Une chaîne lexicale relie les différentes occurrences d'un mot du texte, et éventuellement les mots sémantiquement proches ou les références, si ces occurrences ne sont pas séparées par une distance supérieure à un seuil appelé *hiatus*. Ces chaînes permettent de prendre en compte, non seulement la répétition des mots dans le texte, mais également la plus ou moins grande localité de ces répétitions dans le texte. Morris et Hirst (Morris and Hirst, 1991) ont été les premiers à proposer la prise en compte de ces chaînes pour la segmentation thématique. Ils utilisent, pour cela, des chaînes calculées à partir des occurrences des mots ainsi que des mots qui leur sont sémantiquement reliés dans un thésaurus. Le poids associé à chacune des chaînes dépend de trois éléments : le nombre de répétitions présentes dans la chaîne, sa densité et sa longueur. Ce travail montre que les débuts et fins de chaînes lexicales sont corrélés avec la structure thématique d'un texte (*cf.* Figure 4.1). Dans (Stokes et al., 2002), les chaînes lexicales sont obtenues à partir de mots répétés mais également à partir de mots liés sémantiquement dans WordNet. Un nouvel élément est ajouté à une chaîne s'il est en relation avec au moins un mot de la chaîne. De plus, si un mot peut être en relation avec plusieurs chaînes, il est ajouté à la chaîne la plus récemment mise à jour. Pour définir les frontières thématiques, les auteurs associent à chaque frontière potentielle un score qui correspond au produit entre le nombre de chaînes se terminant à la phrase n par le nombre de chaînes commençant à la phrase $n+1$. Finalement, dans (Sitbon and Bellot, 2005) les auteurs associent à chaque position du texte une valeur qui correspond à la similarité entre le vecteur représentant les poids des chaînes lexicales actives n phrases avant et n phrases après la frontière potentielle. La pondération utilisée prend en compte la compacité des chaînes lexicales, c'est-à-dire le *ratio* entre la taille des chaînes lexicales et le nombre d'occurrences de mots qu'elles contiennent.

La segmentation thématique fondée sur le critère de cohésion lexicale peut également être obtenue par le biais d'une technique de classification. Dans (Yaari, 1997), Yaari utilise une méthode de *clustering* agglomératif hiérarchique pour extraire automatiquement une structure thématique d'un texte descriptif. Cette technique consiste, dans un premier temps, à découper le texte en segments élémentaires – ici les paragraphes du texte – puis une phase de fusion est appliquée itérativement. Lors de cette phase, les deux segments consécutifs les plus similaires sont fusionnés. La phase de fusion s'arrête lorsqu'il ne reste plus qu'un segment. La similarité de deux segments se calcule grâce à une mesure cosinus appliquée sur des vecteurs de mots pleins *stemmés* pondérés par un poids *tf-idf*. De cette étape de fusion, l'auteur obtient un dendrogramme (*cf.* Figure 4.2) dans lequel les feuilles représentent les paragraphes du texte et les nœuds les fusions des segments. À partir de ce dendrogramme, Yaari applique deux règles, *le col* et *la falaise*, pour obtenir une segmentation linéaire.

Dans (Bellot and El-Bèze, 2001), les auteurs classent les phrases d'un document afin de regrouper les individus proches les uns des autres, c'est-à-dire les phrases possédant beaucoup de mots en commun. La segmentation est ensuite effectuée en supposant que deux phrases issues de la même classe partagent la même thématique. Au contraire, si deux phrases se trouvent dans deux classes différentes, il est probable qu'elles abordent des sujets différents. La segmentation thématique consiste donc à définir une frontière thématique entre deux phrases consécutives si elles appartiennent à deux classes différentes.

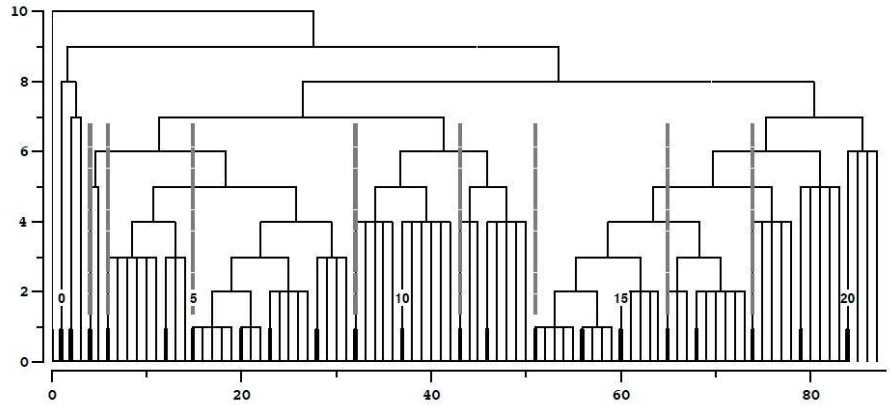


FIG. 4.2 – Dendrogramme pour l'article *Stargazers* (Hearst, 1994). L'échelle sous l'axe des abscisses représente le numéro des phrases et celle au-dessus de l'axe le numéro des paragraphes. Les lignes grises verticales correspondent aux frontières proposées pour la segmentation linéaire. L'axe des ordonnées correspond à la profondeur du chemin entre un nœud et les feuilles du dendrogramme.

4.2.1.2 Méthodes globales basées sur la mesure de la cohésion lexicale

Si les techniques de segmentation thématique locales définissent les frontières thématiques en comparant les régions du document à segmenter dans un voisinage restreint, les méthodes globales opèrent une comparaison entre toutes les régions du document et cherchent à maximiser la valeur de la cohésion lexicale à l'intérieur de chacun des segments définis par les frontières.

Une grande famille des méthodes de segmentation thématique globales se base sur une représentation matricielle du document (*cf.* Figure 4.3). Cette représentation matricielle a été proposée pour la première fois dans l'algorithme DOTPLOT définit par Reynar dans (Reynar, 1994). Dans ce travail, l'auteur commence par tracer les points correspondant aux répétitions lexicales au sein du document à segmenter. Pour chaque mot w plein et lemmatisé du texte apparaissant aux positions x et y dans le document, les quatre points correspondant au produit cartésien de l'ensemble contenant ces positions avec lui-même (c'est-à-dire les points (x, x), (x, y), (y, x) et (y, y)) sont tracés sur le graphe. Grâce à cette représentation, des régions plus denses, correspondant aux sections du document contenant des répétitions de vocabulaire, sont clairement visibles sur le graphe (Figure 4.3). Le principe de l'algorithme de segmentation consiste alors à maximiser la densité de ces régions, c'est-à-dire à minimiser la densité des régions extérieures à ces parties contenant une forte répétition de vocabulaire. L'algorithme mis en place est un algorithme de programmation dynamique qui va ajouter des frontières thématiques (uniquement à la fin des phrases du texte) tant que cette densité extérieure augmente ou jusqu'à ce qu'un nombre de frontières fixées *a priori* soit atteint. Dans (Choi, 2000a), Choi propose une amélioration de cette technique dans son algorithme C99. Il utilise pour cela une représentation vectorielle des phrases et mesure une similarité entre tous les vecteurs qu'il combine avec un système de classement, le calcul de la similarité sur de petits segments n'étant, selon lui, pas suffisamment fiable. Cette modification lui permet d'obtenir des segmentations de meilleure qualité comparées aux résultats fournis par DOT-

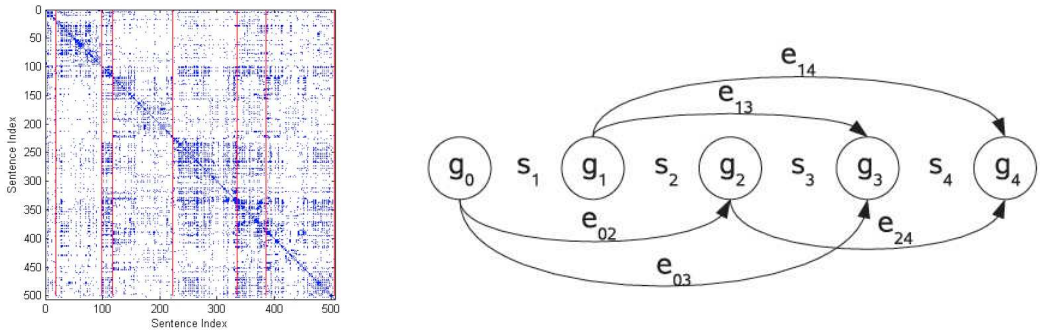
PLOT, surtout sur des documents composés de petits segments. Les auteurs de (Kehagias et al., 2003) suivent la même représentation matricielle du texte, dans laquelle le texte est représenté par une matrice T x T, avec T le nombre de phrases du texte. Une valeur de la matrice $T_{s,t}$ est égale à 1 si les phrases s et t ont un mot en commun et 0 sinon. Ils ajoutent une information *a priori* sur la longueur des segments à retourner. Dans (Ji and Zha, 2003), les auteurs utilisent une technique de segmentation d'images fixes, l'*anisotropic diffusion*, sur la représentation matricielle du texte afin de rehausser la cohésion sémantique des groupes de phrases thématiquement homogènes et de rendre plus nettes les frontières thématiques. Ils proposent également de modifier la méthode de programmation dynamique afin de pouvoir donner à l'utilisateur la possibilité de choisir le nombre de segments retournés et ainsi définir la granularité de la segmentation.

Toutes les méthodes de segmentation globales ne se basent pas sur une représentation matricielle. (Malioutov and Barzilay, 2006) et (Utiyama and Isahara, 2001), par exemple, représentent le document à segmenter à l'aide d'un graphe dans lequel les nœuds correspondent aux frontières potentielles et les arcs aux segments thématiques (*cf.* Figure 4.3-(b)). Dans (Malioutov and Barzilay, 2006), les auteurs voient la tâche de segmentation comme une tâche de partitionnement de graphe, cherchant à minimiser le critère *normalized cut*. Pour cela, ils mettent en place une méthode qui consiste à découper le graphe en k classes, qui possèdent à la fois un contenu très homogène et qui sont très différentes les unes des autres, tout en minimisant le coût de partitionnement correspondant à la somme des coûts associés à chacun des arcs reliant les k classes. Le coût des arcs est calculé grâce à une mesure de similarité (cosinus) appliquée sur des vecteurs de mots pondérés par des scores *tf-idf*. Dans (Utiyama and Isahara, 2001), les auteurs construisent une segmentation thématique du document en retrouvant le meilleur chemin dans un graphe valué. Le coût associé à chacun des arcs du graphe correspond à une mesure de la cohésion lexicale au sein du segment qu'il représente, cohésion lexicale fondée sur le calcul d'une probabilité généralisée. La valeur de la cohésion lexicale d'un segment S_i est ainsi vue comme la mesure de la capacité d'un modèle de langue Δ_i – c'est-à-dire une distribution de probabilités – appris sur le segment S_i à prédire les mots contenus dans le segment. (Misra and Yvon, 2010) propose une extension de cet algorithme qui permet de fournir, en plus d'une segmentation thématique, une coloration thématique des segments définis. Pour cela, ils utilisent un modèle thématique probabiliste, l'Allocation Dirichlet Latente (LDA). Leur technique consiste à détecter, dans un premier temps, les thèmes latents au sein d'un document. L'algorithme de segmentation de (Utiyama and Isahara, 2001) est ensuite modifié en associant à chaque arc un vecteur contenant la probabilité que le segment correspondant à l'arc soit lié à chacun des thèmes latents détectés, probabilité fournie grâce à la LDA.

4.2.2 Évaluation de la segmentation thématique

Afin d'estimer la qualité d'une segmentation thématique produite par une méthode automatique, la technique d'évaluation consiste généralement à comparer cette segmentation à une segmentation de référence grâce à une métrique. Si cela peut sembler simple dans les faits, cette méthode d'évaluation pose, en réalité, deux problèmes majeurs.

Premièrement, la création d'une segmentation de référence n'est pas une tâche simple à mettre en place. Elle est coûteuse en temps, chaque expert devant saisir l'organisation interne des documents, ce qui nécessite une lecture attentive des données à segmenter. De plus, la création de segmentation de référence se heurte au problème de l'accord inter-annotateurs et



(a) (b)

FIG. 4.3 – Représentation matricielle (a) et graphique (b) des documents à segmenter.

de la granularité. En effet, comme nous l'avons mentionné dans la section 4.1, l'identification d'un thème est très subjective et certains annotateurs peuvent détecter un changement de thématique là où d'autres perçoivent une continuité. Pour pallier ce problème, certains auteurs testent leurs méthodes sur des textes pré-formatés. Les techniques de segmentation sont évaluées sur leur capacité à retrouver le formatage original des documents, c'est-à-dire le découpage du texte en paragraphes, qui sont considérés comme des unités thématiquement homogènes. Un autre type de corpus permettant d'évaluer des méthodes de segmentation thématique est celui proposé par Choi dans (Choi, 2000a). Ce corpus est composé de 700 documents créés à partir de la concaténation de 10 portions de texte, chacune d'entre elles correspondant aux n premières phrases (n variant de 3 à 11) d'articles sélectionnés aléatoirement dans le corpus de Brown. Grâce à ce corpus, Choi permet aux différents auteurs de comparer leur méthode de segmentation thématique avec celles déjà existantes dans la littérature sur la base d'un corpus de test commun. Cependant, ce corpus est extrêmement artificiel et les changements thématiques créés par la concaténation d'articles sont beaucoup plus violents que ceux que l'on trouve dans des documents classiques.

Afin d'évaluer notre méthode de segmentation thématique linéaire, nous avons défini une segmentation de référence en posant une frontière thématique à chaque nouveau reportage (dans le sens évoqué en section 4.1.2). Cette séparation est généralement très marquée (souvent associée à un retour plateau) et laisse peu de place à l'interprétation.

La seconde difficulté rencontrée lors de l'évaluation d'une méthode de segmentation thématique réside dans le choix de la métrique d'évaluation. Parmi les mesures d'évaluation actuelles, on retrouve couramment les mesures de *rappel* et de *précision* empruntées à la recherche d'information. Dans le cadre de l'évaluation de la segmentation thématique, ces deux mesures se définissent de la façon suivante :

$$\text{Rappel} = \frac{|H \cap R|}{|R|} \quad , \tag{4.1}$$

$$\text{Précision} = \frac{|H \cap R|}{|H|} \quad , \tag{4.2}$$

avec $|H|$ (*resp.* $|R|$) le nombre de frontières contenues dans la segmentation hypothèse (*resp.* de référence). Ainsi, le rappel correspond à la proportion de frontières de référence à avoir été détectées par la méthode à évaluer, et la précision représente le *ratio* de frontières produites appartenant à la segmentation de référence. Le rappel et la précision peuvent être combinés

dans une seule mesure, notée F_ζ, associée à un paramètre ζ permettant de favoriser le rappel ou la précision dans le calcul de la mesure.

$$\mathrm{F}_\zeta = (1+\zeta^2) * \frac{(\text{Précision} * \text{Rappel})}{(\zeta^2 * \text{Précision}) + \text{Rappel}} \tag{4.3}$$

Dans (Beeferman et al., 1997), Beeferman *et al.* mettent en évidence deux limites majeures à ces mesures. Premièrement, les deux critères de rappel et de précision ne permettent pas de comparer clairement deux segmentations proposant des nombres de frontières différents car il n'est pas pas possible de sélectionner, au sein d'une segmentation, un nombre restreint de frontières à évaluer. Deuxièmement, ces mesures ne prennent en compte que les correspondances exactes entre la segmentation de référence et celle à évaluer. Un décalage d'une frontière d'une phrase ou deux est alors aussi pénalisant que l'absence de frontière.

Pour dépasser ces limites, Beeferman *et al.* proposent une mesure d'évaluation alternative, la mesure P_k. Cette mesure d'évaluation repose sur le principe d'une fenêtre glissante de taille k[4] parcourant la segmentation de référence R et la segmentation hypothétique H proposée par le système à évaluer. P_k consiste à évaluer la similarité entre les deux segmentations au sein de la fenêtre. Cette similarité est calculée grâce à la fonction f qui mesure le désaccord entre les deux segmentations. Il existe un désaccord entre la segmentation de référence et la segmentation hypothèse si les deux extrémités de la fenêtre appartiennent au même segment dans un cas et pas dans l'autre. Formellement, la mesure P_k est définie par :

$$\mathrm{P}_k = \frac{1}{N-k} \sum_{i=1}^{N-k} f(f(r_i, r_{i+k}), f(h_i, h_{i+k})) \ , \tag{4.4}$$

avec N le nombre d'atomes dans le texte, r_i l'indice du segment contenant le i^e atome dans la référence et h_i l'indice du segment contenant le i^e atome dans l'hypothèse. La fonction f est égale à 1 si ses deux arguments sont égaux et à 0 sinon.

Après avoir procédé à une analyse de cette mesure, Pevzner et Hearst ont néanmoins identifié un certain nombre de biais (Pevzner and Hearst, 2002). Tout d'abord, l'absence d'une frontière à une position donnée (frontière manquante) est plus pénalisée que l'insertion d'une frontière n'existant pas dans la référence (frontière abusive). De plus, les décalages de frontières sont, selon eux, trop pénalisés. Finalement, la mesure est sensible aux variations de taille des segments. Par conséquent, Pevzner et Hearst ont proposé une nouvelle mesure d'évaluation qui s'inspire de P_k, la mesure *WindowDiff*. Cette mesure considère le nombre de frontières entre deux phrases séparées par une distance k lors du calcul du degré de similarité entre deux segmentations. La fonction f prend ainsi la valeur du nombre de frontières existant entre les deux bords de la fenêtre plutôt qu'une valeur binaire. Dans (Pevzner and Hearst, 2002), les auteurs ont montré que cette mesure permettait de pallier les limites relatives à P_k. Plus précisément, elle permet de gagner en stabilité face aux variations de taille des segments et se révèle aussi sévère avec les frontières manquantes qu'avec les frontières abusives.

Cependant un biais existant dans la définition de la mesure P_k subsiste malgré la modification proposée dans (Pevzner and Hearst, 2002) : le fonctionnement basé sur une fenêtre glissante fait que les deux mesures donnent moins de poids, lors de l'évaluation, aux frontières proches du début et de la fin du document. De plus, (Carroll, 2010) considère que la mesure

[4]La valeur optimale du paramètre k a été défini comme étant la moitié de la taille moyenne des segments dans la segmentation de référence, taille calculée en terme de nombre d'atomes présents dans les segments, c'est-à-dire de mots ou de phrases (Beeferman et al., 1997).

TAB. 4.1 – Performances des algorithmes de segmentation thématique

	Journaux télévisés			*Sept à Huit*		
Algorithmes	précision (P)	rappel (R)	mesure F_1	précision (P)	rappel (R)	mesure F_1
UI	**57,6**	**61,4**	**59,4**	**60**	50,6	54,9
DOTPLOT	36,4	36,4	36,4	49,5	49,5	49,5
C99	50,2	50,2	50,2	57,4	**57,4**	**57,4**
TEXTTILING	42,0	36,0	38,8	26,0	21,3	23,4

WindowDiff pénalise les faux positifs et les faux négatifs de manière égale, ce qui la conduit à favoriser les segmentations contenant peu de frontières.

Afin d'évaluer nos algorithmes de segmentation thématique thématique linéaire, nous avons donc choisi d'utiliser les mesures de *rappel* et de *précision*, ainsi que leur combinaison par la mesure F_1, bien qu'elles ne soient pas définies originellement pour l'évaluation de la segmentation thématique. En effet, ces mesures ne présentent pas le risque d'être pénalisées par une grande variation dans la longueur des segments comme P_k, ou de favoriser les segmentations contenant peu de frontières. Afin de pallier les biais mis en avant par Beeferman *et al.* à propos de ces critères d'évaluation, nous mettons en place un alignement entre notre segmentation de référence et notre segmentation hypothèse avant de calculer les valeurs de rappel et de précision. De plus, pour limiter le problème de la mise en correspondance exacte des frontières, nous autorisons nos frontières hypothèses à être éloignées de 10 secondes des frontières de référence. Si cette valeur peut paraître élevée, elle s'explique par le fait que nous travaillons sur des données télévisuelles et que notre méthode de segmentation se base sur la parole prononcée. Or, il existe souvent dans les émissions télévisées un décalage entre la fin de la parole prononcée et la fin d'un reportage.

4.3 Approche retenue

La phase de segmentation thématique de nos émissions télévisuelles constituant la base de nos méthodes de structuration, le choix de l'approche utilisée pour réaliser cette segmentation est crucial. En effet, une comparaison menée dans (Sitbon and Bellot, 2004) entre plusieurs algorithmes de segmentation thématique sur des corpora ayant des caractéristiques très différentes montre que la taille des documents segmentés, leurs types et la variation de la taille des segments qui les composent sont autant d'éléments influençant les performances de l'algorithme. À titre d'exemple, le tableau 4.1[5] résume les valeurs de rappel, précision et de la mesure F_1[6], pour différents algorithmes de l'état de l'art, présentés dans la section 4.2.1, appliqués sur nos corpora de journaux télévisés et d'émissions *Sept à Huit*.

Ce tableau montre clairement deux choses. D'une part, l'algorithme TEXTTILING, s'il est efficace sur des corpora composés de segments de tailles peu variables, fournit des résultats peu satisfaisants sur nos corpora. Cette observation s'explique par le fait que les

[5]Les algorithmes utilisés pour cette comparaison ont été réimplémentés et adaptés au français par Sitbon *et al.* (Sitbon and Bellot, 2004). Les résultats pour les algorithmes C99 et DOTPLOT ont été obtenus avec des versions prennant en compte le nombre de segments attendu en sortie, ce qui n'est pas le cas pour les algorithmes UI et TEXTTILING.

[6]Les valeurs de rappel, précision et de mesure F_1 ont été obtenues pour les segmentations dans lesquelles le nombre de segments est le plus proche du nombre de segments contenu dans la référence.

méthodes locales, comme TextTiling, qui calculent la valeur de la cohésion lexicale associée à une frontière potentielle en se basant sur l'étude d'un voisinage de taille fixe, sont très sensibles aux variations de la taille des segments composant le document à segmenter. D'autre part, ce tableau nous indique que les résultats de l'algorithme de Utiyama et Isahara, UI ,sont meilleurs que ceux des trois autres algorithmes sur le corpus de journaux télévisés et en deuxième position sur le corpus *Sept à Huit*. Si les résultats sont un peu plus faibles que ceux de C99 pour le second corpus, l'algorithme possède selon nous l'avantage de ne pas nécessiter d'information *a priori* concernant le nombre de segments attendu. Remarque:Les valeurs proposées dans le tableau, pour *Sept à Huit*, ont, en fait, été calculées sur un corpus composé des émissions *Sept à Huit* et *Envoyé Spécial*. Les performances sur les seuls *Sept à Huit* sont sans doute un peu différentes. À calculer ! !

Nous avons donc sélectionné l'algorithme de Utiyama et Isahara comme point de départ pour mener à bien notre tâche de segmentation thématique, pour sa stabilité vis-à-vis de la variabilité de la taille des segments, parce qu'il est l'un des meilleurs algorithmes de l'état de l'art qui ne fait aucune supposition sur le type de données à traiter, mais également pour une troisième raison liée à la nécessité d'adapter cet algorithme à nos données particulières. En effet, il présente l'avantage d'autoriser l'ajout d'informations additionnelles permettant d'améliorer la qualité de la segmentation.

Cependant, si nous avons choisi d'utiliser une méthode de segmentation globale, basée sur la mesure de la cohésion lexicale, pour gérer la variabilité de la taille des segments, la détection de la rupture de cohésion mise en œuvre dans les méthodes de segmentation locales nous semble être une information intéressante à prendre en compte. En effet, une méthode combinant détection de rupture et mesure de cohésion lexicale doit permettre selon nous d'obtenir une segmentation dans laquelle les segments résultants sont à la fois très homogènes et très différents les uns des autres, comme dans les travaux proposés par (Malioutov and Barzilay, 2006). Par conséquent, si les méthodes classiques de segmentation thématique se basent soit sur la mesure de la cohésion lexicale, soit sur la détection de rupture de ce critère, l'originalité de notre approche de segmentation consiste à utiliser conjointement ces deux informations.

Dans cette section, nous présentons tout d'abord l'algorithme de segmentation tel qu'il a été proposé par Utiyama et Isahara. Puis nous décrivons, dans la sous-section suivante, l'extraction des informations de rupture de la cohésion lexicale, obtenues grâce à une méthode développée par Claveau *et al.* (Claveau and Lefèvre, 2011a), ainsi que la façon dont ces informations sont intégrées dans l'algorithme UI.

4.3.1 Segmentation thématique basée sur la maximisation du critère de cohésion lexicale

L'approche de segmentation thématique retenue repose sur la notion de la cohésion lexicale et donc sur la répétition du vocabulaire présent dans la transcription. Afin de repérer plus aisément ces répétitions, un certain nombre de prétraitements classiques ont été appliqués sur les transcriptions. Dans cette section, nous allons tout d'abord décrire ces pré-traitements, puis nous présentons la méthode de calcul de la cohésion lexicale telle qu'elle est mise en place dans l'algorithme de Utiyama et Isahara. Finalement, nous expliquons le fonctionnement de l'algorithme de segmentation.

4.3.1.1 Prétraitements

La répétition des mots présents dans les transcriptions étant la base de l'algorithme de segmentation, une phase préalable de filtrage des mots utilisés pour calculer ces répétitions est nécessaire au bon fonctionnement de l'algorithme. En effet, certains mots apparaissent très fréquemment dans les transcriptions sans que leur valeur informative soit importante. C'est le cas, par exemple, des mots vides et de certains verbes employés la plupart du temps comme verbes modaux ou comme auxiliaires (*être, avoir, falloir, etc.*). Afin de ne retenir que les mots les plus informatifs pour la segmentation thématique, nous les avons étiquetés grâce à l'outil TreeTagger (Schmid, 1994), et n'avons considéré lors de la segmentation que les noms, verbes, autres que les verbes modaux et les auxiliaires, et adjectifs. De plus, afin de favoriser la détection des répétitions, les mots ont été lemmatisés, toujours à l'aide de TreeTagger, afin de les ramener à une forme unique.

4.3.1.2 Mesure de la cohésion lexicale

La méthode présentée dans (Utiyama and Isahara, 2001) repose sur le calcul d'une probabilité généralisée pour mesurer la cohésion lexicale. La valeur de la cohésion lexicale d'un segment S_i est vue comme la mesure de la capacité d'un modèle de langue Δ_i appris sur le segment S_i à prédire les mots contenus dans le segment. Cette définition de la cohésion lexicale nécessite de calculer, dans un premier temps, un modèle de langue Δ_i pour chaque segment S_i du texte à segmenter, puis de déterminer la probabilité généralisée des mots du segment S_i, traduisant la capacité du modèle de langue Δ_i à prédire les mots de S_i.

Modèle de langue Le modèle de langue utilisé pour le calcul de la cohésion lexicale est un modèle *unigramme*[7], qui détermine la probabilité d'apparition de chaque mot plein au sein du texte. Lors de l'estimation du modèle de langue Δ_i d'un segment S_i, on évalue la probabilité d'apparition de chacun des mots du vocabulaire du texte dans le segment S_i. Afin d'éviter que toute la masse de probabilité soit attribuée aux seuls mots apparaissant dans le segment, on applique un lissage à ce modèle de langue dans le but de redistribuer une partie des probabilités aux mots non observés – le nombre de mots observés dans le segment étant relativement petit au regard du nombre de mots dans le texte. Le calcul du modèle de langue du segment S_i se formalise par

$$\Delta_i = \{P_i(u) = \frac{C_i(u) + 1}{z_i},\ \forall\ u \in V_K\}\ , \tag{4.5}$$

avec V_K le vocabulaire du texte, contenant K mots distincts, et $C_i(u)$ le compte du mot u, correspondant à son nombre d'occurrences dans le segment S_i. La distribution de probabilités est lissée en incrémentant le compte de chacun des mots de 1. On a donc $z_i = K + \sum_{u \in V_K} C_i(u)$.

Calcul de la probabilité généralisée La seconde étape du calcul de la cohésion lexicale d'un segment consiste à évaluer une probabilité traduisant à quel point le modèle de langue

[7] L'utilisation d'un modèle de langue *unigramme* s'explique ici par le fait que ce modèle sert à représenter la répétition de vocabulaire au sein de nos données. De fait, si l'utilisation d'un modèle de langue *bigramme* (ou d'ordre supérieur) est plus adaptée pour une tâche de modélisation du langage, un modèle de langue *unigramme* nous permet de représenter la répétition de termes simples dans nos segments.

Δ_i permet d'expliquer les mots contenus dans le segment S_i, soit

$$\ln P[S_i|\Delta_i] = \sum_{j=1}^{n_i} \ln P[w_j^i; \Delta_i]) \ , \tag{4.6}$$

avec n_i le nombre de mots dans le segment et w_j^i le j^e mot du segment. Intuitivement cette probabilité favorise les segments les plus cohérents lexicalement, puisque sa valeur est plus importante lorsque les mots apparaissent plusieurs fois au sein du segment et qu'elle atteint sa valeur minimale lorsque tous les mots du segment sont différents.

4.3.1.3 Algorithme de segmentation thématique

L'algorithme de segmentation thématique développé par Utiyama et Isahara consiste à rechercher la segmentation thématique produisant les segments les plus cohérents d'un point de vue lexical, tout en respectant une distribution *a priori* de la longueur des segments. Dans un cadre probabiliste, l'objectif de l'algorithme est de trouver la segmentation d'une séquence de l unités (dans notre cas des groupes de souffle) $W = W_1^l$ parmi toutes les segmentations possibles,

$$\hat{S} = \underset{S_1^m}{\operatorname{argmax}} P[W|S] \ P[S] \ . \tag{4.7}$$

En supposant que $P[S_1^m] = n^{-m}$, avec n le nombre de mots du texte et m le nombre de segments, et en supposant que les segments sont indépendants, la probabilité d'un texte W pour une segmentation $S = S_1^m$ est donnée par

$$\hat{S} = \underset{S_1^m}{\operatorname{argmax}} \sum_{i=1}^{m} \Big(\ln(P[W_{a_i}^{b_i}|S_i]) - \alpha \ln(n) \Big) \ , \tag{4.8}$$

avec $P[W_{a_i}^{b_i}|S_i]$ la probabilité généralisée de la séquence d'unités correspondant au segment S_i tel que définie par (4.6). Le paramètre α permet de faire un compromis entre la cohésion lexicale et la longueur des segments retournés.

4.3.2 Combinaison de la mesure et de la détection de rupture de cohésion lexicale

L'algorithme de segmentation thématique, tel qu'il a été défini dans la sous-section précédente, a pour objectif de produire une segmentation maximisant une mesure de la cohésion lexicale. Afin d'obtenir une segmentation contenant des segments à la fois très cohérents d'un point de vue lexical et très différents les uns des autres, nous proposons intégrer à cet algorithme des informations de rupture de la cohésion lexicale au sein du document.

Afin de combiner la mesure de la cohésion lexicale et des informations de rupture de cette cohésion, nous associons[8] dans un premier temps des scores à chaque frontière potentielle appartenant à nos documents, c'est-à-dire chaque frontière entre deux groupes de souffle. Ces scores sont obtenus grâce à l'algorithme de Claveau *et al.* (Claveau and Lefèvre, 2011a) décrit dans la section 4.3.2.1. Ils sont ensuite introduits dans l'algorithme de Utiyama et Isahara et combinés à la mesure de la cohésion lexicale dans un cadre probabiliste, comme présenté en section 4.3.2.2.

[8] Ce travail a été réalisé dans le cadre du stage de Master 1 Informatique de Clémentine Maurice, co-encadrée par Camille Guinaudeau, Vincent Claveau et Guillaume Gravier.

4.3.2.1 Détection de rupture de la cohésion lexicale

L'algorithme de segmentation développé par Claveau et Lefèvre (Claveau and Lefèvre, 2011a) repose sur un principe de fenêtre glissante. Dans cet algorithme, chaque frontière potentielle est associée à un score traduisant la proximité sémantique entre les n groupes de souffle situés avant et les n groupes de souffle situés après cette frontière.

Comparativement aux autres méthodes classiques de segmentation thématique locale, la technique proposée, présente l'avantage de gérer la présence de synonymes et une faible répétition du vocabulaire grâce à une comparaison indirecte des éléments du texte. Cette comparaison indirecte, qui consiste à mesurer la similarité entre les éléments du texte et des documents-pivots plutôt qu'entre eux, repose sur l'idée que si deux documents d_1 et d_2 sont sémantiquement proches d'un troisième document d_3 alors d_1 et d_2 partagent un lien de sens. Ainsi, la valeur de la cohésion lexicale associée à une frontière potentielle correspond à la similarité entre les vecteurs caractéristiques des n groupes de souffle situés avant la frontière et les n groupes de souffle situés après, ces vecteurs étant composés de m scores qui traduisent la proximité sémantique entre les groupes de souffle et m documents-pivots.

Plus formellement, la valeur de la cohésion lexicale associée à une frontière potentielle i se calcule par :

$$\bigtriangledown(i) = L_2(\text{Vect}(Pred(i-n), \varphi, \sqrt{TF}/L_2)), \text{Vect}(Succ(i+n), \varphi, \sqrt{TF}/L_2) \ , \qquad (4.9)$$

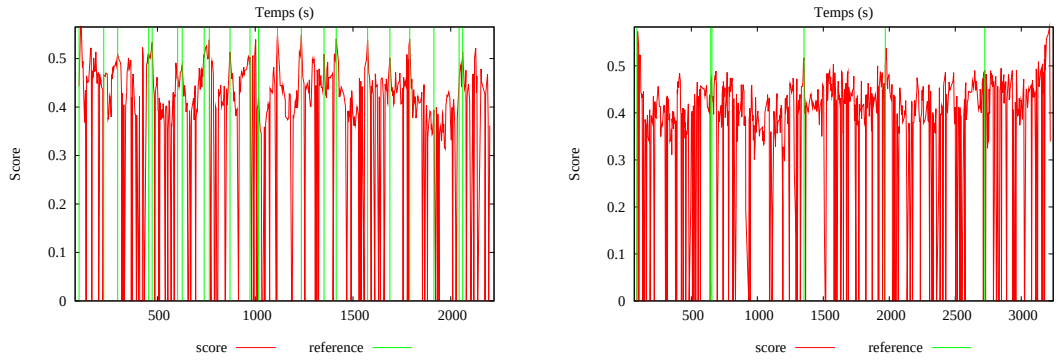
avec $\text{Vect}(Pred(i-n), \varphi, \sqrt{TF}/L_2)$ le vecteur représentatif des groupes de souffle précédents la frontière potentielle i, composés des scores comparant les mots apparaissant dans ces groupes de souffle et l'ensemble des documents φ. Ces scores sont obtenus en calculant une mesure de similarité L_2 entre les vecteurs de mots pondérés par des scores $\sqrt{TF}$, avec TF le nombre d'occurrences d'un mot. L'ensemble φ est constitué de m segments d'émissions télévisées extraits aléatoirement des données à segmenter.

Si dans (Claveau and Lefèvre, 2011a) les auteurs réalisent ensuite une série de *post-processings*, basée sur une technique de segmentation d'image et prenant en compte des connaissances spécifiques aux données traitées pour extraire des frontières thématiques (voir l'article pour plus de détail), nous n'utilisons que les scores, normalisés, dans le cadre de la combinaison d'informations de rupture et de mesure de la cohésion lexicale.

La figure 4.4 présente les valeurs des scores $\bigtriangledown(i)$ pour chaque frontière potentielle i de deux documents extraits de nos corpora de journaux télévisés (JT) et d'émissions de reportages *Sept à Huit*. Nous pouvons constater, grâce à ce graphique, que, pour le journal télévisé, les valeurs de score les plus importantes correspondent globalement aux frontières thématiques de référence (lignes verticales vertes). Cette corrélation entre frontières thématiques et valeurs de score élevées n'est en revanche pas observée dans le corpus *Sept à Huit*. On remarque, en effet, dans la partie droite de la figure 4.4 que de nombreux « pics » dans le score $\bigtriangledown(i)$ (courbe rouge) apparaissent au milieu de segments thématiques.

4.3.2.2 Introduction des informations de rupture dans l'algorithme de segmentation

Dans (Huet et al., 2008), les auteurs proposent une extension de l'algorithme UI permettant l'introduction de nouvelles informations. Si l'algorithme de segmentation de Utiyama et Isahara peut être vu comme la recherche du meilleur chemin dans un graphe valué, les arcs du graphe représentant les segments thématiques et les nœuds les frontières potentielles, la valeur



corpus *JT* corpus *Sept à Huit*

FIG. 4.4 – Score $\bigtriangledown(i)$ pour chaque frontière potentielle i dans un journal télévisé et dans une émission de reportages *Sept à Huit*.

de la cohésion lexicale de chacun des segments correspond au poids associé à chacun des arcs du graphe. Dans cette représentation graphique du fonctionnement de l'algorithme, l'introduction d'une nouvelle information peut se voir comme le fait d'associer un poids à chacun des nœuds du graphe permettant de favoriser une segmentation passant par un nœud associé à un brusque changement de vocabulaire. Plus formellement, l'objectif de la segmentation, décrit par (4.7) est redéfini de la façon suivante :

$$\hat{S} = \underset{S_1^m}{\operatorname{argmax}}\, P[W|S]\ P[S|R]^{\beta}\ P[S]\ \ , \tag{4.10}$$

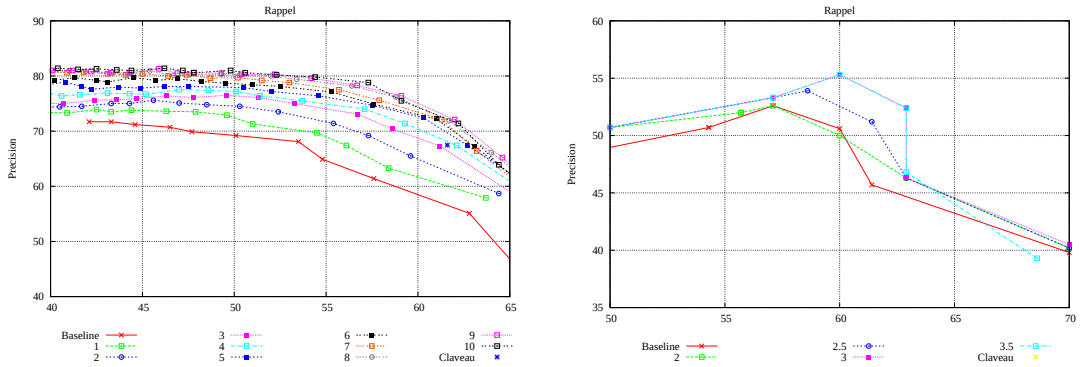
avec $P[S|R]$ la probabilité d'obtenir une segmentation S connaissant les informations de rupture de la cohésion lexicale associées au document à segmenter. Le paramètre β permet de tenir compte des différences de facteur d'échelle entre les différentes probabilités et de donner plus ou moins de poids aux informations de rupture. La valeur du coût d'un segment S_i constitué des groupes de souffle $s_a ... s_b$ devient donc

$$\begin{aligned} C(S_i|W,R) \quad = \quad & -\ log P[W_i|S_i] \\ & -\beta \left[\sum_{j=a}^{b-1} log\ P\ (B_j = \text{« non »}|R_j) + log\ P\ (B_b = \text{« oui »}|R_b) \right] \\ & -\alpha\ log\ P(S_i) \end{aligned} \tag{4.11}$$

avec $P\ (B_j = \text{« non »}|R_j)$ (resp. $P\ (B_b = \text{« oui »}|R_b)$) la probabilité qu'il n'y ait pas de (resp. qu'il y ait une) frontière thématique après le groupe de souffle j (resp. b) connaissant le score de cohésion lexicale associée à la frontière potentielle située après le groupe de souffle b. Cette probabilité est calculée à partir des scores précédemment définis par (4.9) :

$$log\ P\ (B_i = \text{« oui »}|L_i) = log\ \ (\bigtriangledown(i))\ \ . \tag{4.12}$$

Les informations de rupture de cohésion lexicale ont été prises en compte pour la segmentation des journaux télévisés et des émissions de reportages *Sept à Huit*. La figure 4.5 décrit les performances de l'algorithme de segmentation ainsi modifié. Les courbes rappel/précision, présentées dans la figure ont été obtenues en faisant varier le paramètre α permettant d'influer



corpus *JT* corpus *Sept à Huit*

FIG. 4.5 – Courbe rappel/précision sans et avec intégration d'informations de rupture dans l'algorithme de segmentation thématique.

sur le nombre de frontières retournées par l'algorithme. Sur cette figure, les courbes rouges représentent la segmentation obtenue par l'algorithme de segmentation UI, sans ajout d'informations de rupture de la cohésion lexicale. Les autres courbes correspondent à différentes valeurs du paramètre β, donnant plus ou moins d'importance aux informations de rupture, variant de 1 à 10 pour les journaux télévisés et de 2 à 3,5 pour les émissions *Sept à Huit*.

Concernant les journaux télévisés, nous pouvons constater que la combinaison de la *mesure de la cohésion lexicale* et de la *détection de rupture* de cette cohésion permet d'améliorer la qualité de la segmentation obtenue. Pour un paramètre β supérieur à 4 les segmentations proposées sont, en effet, meilleures que la segmentation thématique globale et que la segmentation locale (étoile bleue) définie dans (Claveau and Lefèvre, 2011a). Le gain observé augmente avec la valeur du paramètre β et l'on constate une stagnation de cette amélioration lorsque β est égal à 10.

Les résultats obtenus pour les émissions de reportages *Sept à Huit* sont, en revanche, moins intéressants. En effet, bien que la combinaison des deux indices conduise à une amélioration de la qualité de la segmentation par rapport à celle fournie par la technique globale, ce gain est moins important que celui observé pour les journaux télévisés. De plus, l'intégration d'informations de rupture de la cohésion lexicale dans l'algorithme de segmentation ne permet pas d'outrepasser les performances du système de Claveau et Lefèvre. `Remarque:Le point Claveau n'apparaît pas dans la figure (il est au dessus) mais sa valeur a été calculée sur les émissions Sept à Huit et Envoyé Spécial. Je dois donc la recalculer.`

Cette différence de performance entre les deux corpora s'explique par la qualité des scores obtenus, présentés dans la figure 4.4 ; nous avions, en effet, constaté sur ces graphiques que l'augmentation des valeurs du score $\bigtriangledown(i)$ n'était pas toujours corrélée avec les frontières thématiques dans le cas des émissions de reportages. L'influence de la qualité des scores introduits est également perceptible grâce aux valeurs du paramètre β utilisées pour ce corpus, celles-ci étant plus faibles pour les émissions *Sept à Huit* ; une grande valeur de β entraîne une dégradation des performances de l'algorithme de segmentation.

4.4 Bilan du chapitre

Dans ce chapitre, nous avons tout d'abord présenté les différentes méthodes de segmentation thématique existantes qu'elles soient fondées sur la détection de rupture du critère de cohésion lexicale ou sur sa maximisation. Nous avons également proposé dans ce chapitre une approche originale de segmentation qui fusionne ces deux types de techniques, habituellement employées indépendamment l'un de l'autre. Plus précisément, nous avons décrit le principe de fonctionnement de la méthode de segmentation globale qui sert de base à notre segmentation thématique. Puis nous avons exposé la technique employée pour intégrer des informations de rupture au sein de la mesure de la cohésion lexicale. Nous avons pu montrer que cette intégration permettait d'améliorer la qualité de la segmentation thématique, que ce soit pour les journaux télévisés ou les émissions de reportages *Sept à Huit*. Les résultats obtenus pour l'utilisation conjointe des deux techniques de segmentation thématique étant plutôt encourageants, deux pistes nous semble particulièrement intéressantes pour améliorer la qualité de cette combinaison. Tout d'abord les informations de rupture de la cohésion lexicale intégrées dans l'algorithme de segmentation globale n'ayant pas subit de *post-processing*, l'optimisation de ces informations devrait, selon nous, permettre de conduire à une segmentation thématique de meilleure qualité. Cette optimisation n'est cependant pas évidente à mettre en place. En effet, des premiers tests effectués dans le but d'obtenir des informations de rupture moins bruitées ont montré qu'une telle optimisation ne conduisait pas à une meilleure segmentation. Cette constatation s'explique, selon nous, par le fait que la décision finale de placer ou non une frontière thématique est, dans ce cas, trop influencée par l'une ou l'autre des méthodes et non plus par la combinaison des deux. Afin de perfectionner la combinaison des deux techniques, la deuxième piste qui nous semble la plus intéressante consiste à détecter la rupture de la cohésion lexicale et à maximiser la valeur de la cohésion lexicale au sein des segments en même temps dans l'algorithme de Utiyama et Isahara. Ainsi l'algorithme résultant serait capable de proposer une segmentation thématique composée de segments très cohérents et très différents les uns des autres sans qu'il soit nécessaire de calculer par avance les informations de rupture.

Chapitre 5

Adaptation de la cohésion lexicale aux particularités des documents oraux

La méthode de segmentation thématique proposée par Utiyama et Isahara, développée pour du texte écrit, voit ses performances diminuer lorsqu'elle est employée sur des transcriptions de documents oraux. Nous allons donc nous intéresser, dans ce chapitre, à l'adaptation de cet algorithme pour la segmentation de nos émissions télévisuelles. Pour cela, nous proposons différentes techniques visant à rendre le critère de la cohésion lexicale robuste aux spécificités de nos transcriptions. D'une part, ces méthodes, présentées dans la section 5.1, consistent à intégrer des sources d'informations additionnelles, indépendantes du type d'émissions télévisuelles considéré, lors du calcul de la valeur de la cohésion lexicale. Ainsi, les mesures de confiance fournies par le système de reconnaissance de la parole et des relations sémantiques ont pour objectifs respectifs de pallier les erreurs de transcription et de gérer le faible taux de répétitions de vocabulaire. D'autre part, deux techniques d'interpolation de modèles de langue sont utilisées afin de proposer de meilleures estimation des modèles de langue nécessaires au calcul de la cohésion lexicale lorsque les segments considérés sont courts.

De plus, la mesure de la cohésion lexicale proposée par Utiyama et Isahara repose uniquement sur les caractéristiques textuelles des documents. Or, il important de prendre en compte le fait que nos données sont audiovisuelles et qu'elles doivent être, de ce fait, étudiées grâce à un spectre d'indices plus large que la seule répétition de vocabulaire. Afin de tirer parti de la multimodalité de nos données, nous proposons, dans ce chapitre, d'intégrer des informations prosodiques dans le calcul de la cohésion lexicale. L'utilisation de la prosodie, décrite en section 5.2, a pour objectif de donner plus de poids aux mots proéminents dans le discours, ceux-ci étant généralement associés à une valeur informative importante.

Chacune des techniques d'adaptation de la cohésion lexicale à nos données télévisuelles a été testée sur les corpora composés de journaux télévisés et d'émissions de reportages *Sept à Huit*. L'impact des méthodes d'adaptation étant plus ou moins important selon les caractéristiques des deux types d'émissions, l'utilisation de ces deux corpora permet de mettre en évidence le comportement des différents indices utilisés vis-à-vis des particularités de données. L'organisation de ce manuscrit ne respectant pas l'ordre chronologique, les expériences décrites dans le chapitre précédent ont été effectuées après celles exposées dans ce chapitre ; les résultats présentés ici n'intègrent donc pas la prise en compte des informations de rupture dans le calcul de la cohésion lexicale, mais les effets bénéfiques sont *a priori* plutôt combinables.

5.1 Gestion des spécificités des transcriptions automatiques de programmes TV

Nous détaillons ici les trois techniques qui ont été employées pour rendre le critère de cohésion lexicale plus robuste aux spécificités des transcriptions automatiques de vidéos professionnelles : l'intégration de mesures de confiance dans le calcul de la cohésion lexicale, la prise en compte de relations sémantiques et l'utilisation de méthodes d'interpolation de modèles de langue.

5.1.1 Mesures de confiance

La cohésion lexicale constituant la base de notre algorithme de segmentation thématique peut être fortement pénalisée par les erreurs de transcription présentes dans nos documents. Nous avons en effet constaté une perte de 15,7 points en précision (pour une même valeur de rappel de 51,3) entre une segmentation thématique effectuée sur les transcriptions manuelles de notre corpus de JT et une segmentation opérée sur les mêmes documents transcrits automatiquement[1]. Cette dégradation est encore plus marquée sur le corpus d'émissions de reportages *Sept à Huit* puisque la précision passe de 90 (pour une valeur de rappel de 78,3) pour les transcriptions manuelles à 60 pour les transcriptions automatiques. Afin de pallier l'influence de ces erreurs de transcription, nous proposons d'intégrer des mesures de confiance au calcul de la cohésion lexicale. Leur intégration a pour objectif de réduire l'importance des mots mal transcrits lors du calcul de la cohésion lexicale. Si l'utilisation d'un tel indice a montré son intérêt dans un algorithme de segmentation thématique locale (Mohri et al., 2009), basé sur la détection de rupture, les mesures de confiance n'ont encore jamais été prises en compte, à notre connaissance, dans un algorithme fondé sur la mesure de la cohésion lexicale.

L'intégration des mesures de confiance peut être effectuée lors des deux phases du calcul de la cohésion lexicale.

Lorsqu'elles sont prises en compte lors de l'estimation du modèle de langue, le compte $C_i(u)$ d'un mot, c'est-à-dire son nombre d'occurrences, est remplacé par la somme des valeurs des mesures de confiance associées à chacune des occurrences de ce mot :

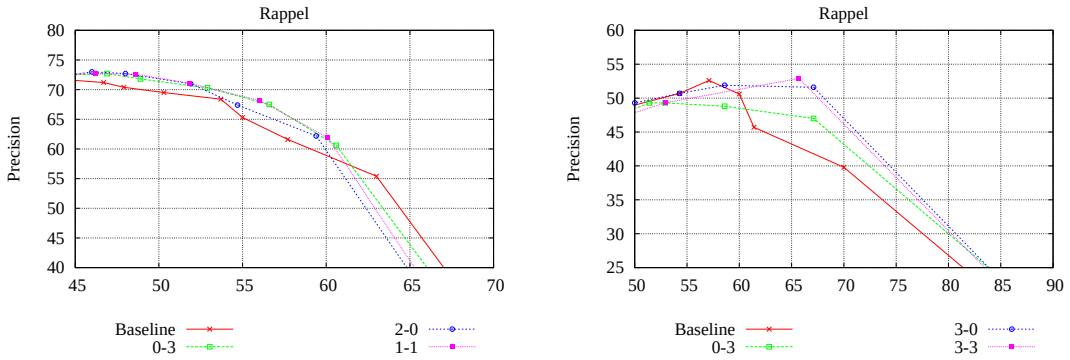
$$C'_i(u) = \sum_{w_j^i = u} c(w_j^i)^{\delta_1} \quad , \tag{5.1}$$

avec $c(w_j^i) \in [0,1]$ la mesure de confiance du j^e mot du segment S_i et δ_1 un paramètre utilisé pour réduire le poids des mots ayant une valeur de mesure de confiance faible. En effet, plus la valeur de δ_1 est élevée, moins l'impact des mots associés à une petite mesure de confiance sera important.

Les mesures de confiance peuvent également être intégrées lors du calcul de la probabilité généralisée. Dans ce cas, la log-probabilité de l'occurrence d'un mot dans un segment est multipliée par la mesure de confiance associée à cette occurrence du mot,

$$\ln P[S_i; \Delta_i] = \sum_{j=1}^{n_i} c(w_j^i)^{\delta_2} \ln P[w_j^i; \Delta_i] \quad , \tag{5.2}$$

[1]La valeur de la précision passe de 91 pour les transcriptions manuelles à 75,3 pour les transcriptions automatiques.



corpus *JT* corpus *Sept à Huit*

FIG. 5.1 – Prise en compte des mesures de confiance

avec δ_2 équivalent à δ_1. L'équation (5.2) permet de réduire la contribution d'un mot de mesure de confiance faible lors du calcul de la cohésion lexicale. Dans ce cas, le modèle de langue Δ_i peut être estimé soit grâce au compte $C_i(u)$, ce qui limite l'intégration des mesures de confiance au calcul de la probabilité généralisée, soit grâce aux comptes modifiés $C'_i(u)$.

La figure 5.1 résume les résultats sur les deux corpora. Sur cette figure, les courbes rouges correspondent à la segmentation obtenue sans prise en compte des mesures de confiance, les autres courbes représentant la segmentation fournie par l'intégration des mesures de confiance avec des valeurs de paramètres δ_1 et δ_2 optimales[2]. Nous pouvons constater, sur cette figure, que l'intégration de mesures de confiance dans le calcul de la cohésion lexicale permet d'améliorer la qualité de la segmentation pour les deux types d'émissions considérés. La valeur de la mesure F_1 est, en effet, augmentée de 2 points pour les journaux télévisés et de 5 points pour les émissions de reportages *Sept à Huit*, augmentations statistiquement significatives d'après le test de Student. Sur cette figure, nous remarquons également que l'augmentation apportée par l'intégration des mesures de confiance est plus importante pour le corpus d'émissions de reportages que pour les journaux télévisés. Cette différence entre les deux corpora est liée, selon nous, à la qualité des transcriptions, le taux d'erreur des transcriptions des émissions *Sept à Huit* étant plus élevé que celui des *JT*.

Afin de valider cette dernière hypothèse, nous avons étudié l'influence des mesures de confiance sur la segmentation thématique de journaux télévisés transcrits avec un système de transcription automatique plus performant que le nôtre sur ce type de corpus. Ce système de reconnaissance automatique de la parole (Gauvain et al., 2002) développé au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) affiche, en effet, un taux d'erreur de VALEUR3 sur notre corpus de journaux télévisés, contre VALEUR pour notre propre système. Les résultats de cette segmentation sont présentés sur la figure 5.2. Sur cette figure, les courbes rouge et orange correspondent aux segmentations obtenues (avec et sans mesures de confiance) sur le corpus de *JT* transcrit avec le système IRENE, tandis que les courbes bleues représentent les résultats pour les transcriptions fournies par le système du LIMSI. À partir de ces courbes, nous constatons que l'influence des mesures de confiance – intégrées lors de l'estimation du modèle de langue et du calcul de la probabilité généralisée avec des paramètres δ_1 et δ_2 égaux à 1 – est moins importante lorsque les transcriptions sont de meilleure qualité. En effet, le gain fourni par les mesures de confiance sur les transcriptions

[2]Pour plus de détail sur l'influence de ces paramètres sur la segmentation thématique se référer à l'annexe B.

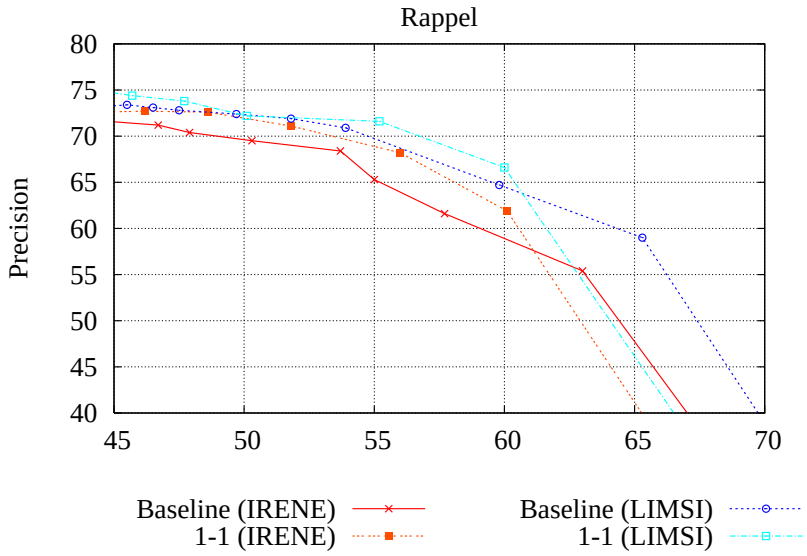
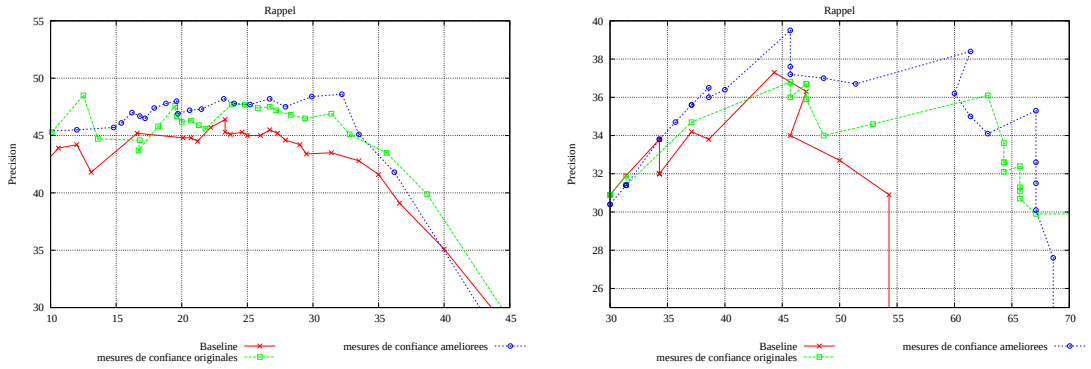


FIG. 5.2 – Influence de la qualité de la transcription sur l'impact des mesures de confiance

du LIMSI est deux fois moins élevé – la valeur de la mesure F_1 passe de 62,15 à 63,13 – que celui observé pour les transcriptions du système IRENE.

Ces deux observations, la différence de performances obtenues sur les deux corpora et l'influence de la qualité des transcriptions des journaux télévisés sur l'impact des mesures de confiance, nous permettent de conclure que les mesures de confiance sont effectivement utiles pour pallier les erreurs de transcription dans notre tâche de segmentation thématique et améliorent la qualité de la segmentation produite.

Dans le chapitre 3, nous avons présenté le travail de (Fayolle et al., 2010) ayant pour objectif d'améliorer la qualité des mesures de confiance, celles fournies par notre système de transcription automatique de la parole n'étant pas toujours fiables. Nous proposons dans une dernière expérience d'évaluer l'impact de la qualité des mesures de confiance sur les performances de la segmentation thématique. Pour cela, nous intégrons les mesures de confiance calculées par Fayolle *et al.* dans le calcul de la cohésion lexicale. Ces mesures de confiance sont prises en compte à la fois lors de l'estimation du modèle de langue et du calcul de la probabilité généralisée avec des valeurs de paramètres δ_1 et δ_2 égales à 1 pour le corpus de journaux télévisés et égales à 3 pour le corpus d'émissions de reportages. Cette expérience nous montre que la qualité accrue des mesures de confiance utilisées améliore le positionnement des frontières thématiques proposées par l'algorithme. En effet, si les résultats ne sont pas meilleurs lorsque les frontières hypothèses sont considérées comme correctes lorsqu'elles sont éloignées de moins de 10 secondes d'une frontière de référence, les performances de l'algorithme sont améliorées lorsque cette différence est ramenée à 2 secondes. La figure 5.3 présente les segmentations obtenues grâce aux mesures de confiance fournies par le système IRENE et celles calculées par (Fayolle et al., 2010). Sur cette figure, nous constatons que les courbes représentant l'utilisation des nouvelles mesures de confiance sont au-dessus de celles correspondant aux résultats fournis par les mesures de confiance classiques, pour les deux types d'émissions. Ainsi, si la qualité des mesures de confiance n'a pas d'impact sur le nombre de frontières correctes retournées par l'algorithme, elles permettent, cependant, d'améliorer leur positionnement.



corpus *JT* corpus *Sept à Huit*

FIG. 5.3 – Influence de la qualité des mesures de confiance sur la segmentation thématique de journaux télévisés et d'émissions de reportages *Sept à Huit*. Les courbes rappel/précision sont calculées en considérant une frontière hypothèse comme correcte lorsqu'elle est éloignée de moins de 2 secondes d'une frontière de référence

5.1.2 Relations sémantiques

Nous proposons également d'adapter le critère de la cohésion lexicale aux particularités des transcriptions automatiques en renforçant la cohésion entre les unités lexicales différentes qui présentent un lien sémantique important par le biais de relations sémantiques. Il a, en effet, été montré que l'utilisation de relations sémantiques permettait d'améliorer la qualité de la segmentation thématique (Ferret, 2006; Bestgen, 2006) lorsque les documents à segmenter étaient caractérisés par une faible répétition de vocabulaire. Notre démarche se distingue de ces travaux de deux façons. Premièrement, nous nous différencions de (Ferret, 2006) par la méthode de segmentation utilisée (opposition entre détection de ruptures et mesure de cohésion). Deuxièmement, contrairement au travail présenté dans (Bestgen, 2006), les relations utilisées sont extraites à partir d'un corpus indépendant de celui fourni en entrée de l'algorithme de segmentation. De plus, ces travaux utilisent les relations sémantiques dans le seul but de pallier la faible répétition du vocabulaire présent dans leurs données. Or nous souhaitons montrer, dans notre thèse, que ces relations sont également utiles pour atténuer l'impact des erreurs de transcription. En effet, contrairement aux mots correctement transcrits, les mots mal reconnus par le système de transcription ont peu de chance d'être liés sémantiquement aux autres mots du segment. De ce fait, lors de l'intégration de relations sémantiques, les mots mal transcrits auront, selon nous, un impact moins important dans le calcul de la cohésion lexicale.

Comme pour les mesures de confiance, l'intégration des relations sémantiques peut être effectuée en modifiant les comptes des mots lors de l'estimation du modèle de langue. Le compte d'un mot, qui reflète normalement le nombre de fois où ce mot apparaît dans un segment, est étendu afin de prendre en compte à la fois le nombre d'occurrences du mot et les occurrences des mots qui lui sont sémantiquement liés. Plus formellement, le compte $C_i(u)$ d'un mot est modifié de la façon suivante :

$$C_i''(u) = C_i(u) + \sum_{j=1, w_j^i \neq u}^{n_i} r(w_j^i, u) \ , \tag{5.3}$$

avec $r(w_j^i, u) \in [0;1]$ la proximité sémantique des mots w_j^i et u, proche de 1 lorsque w_j^i et u partagent un lien sémantique très important. Le calcul de r est décrit en détail dans la section 3.1.2.5.

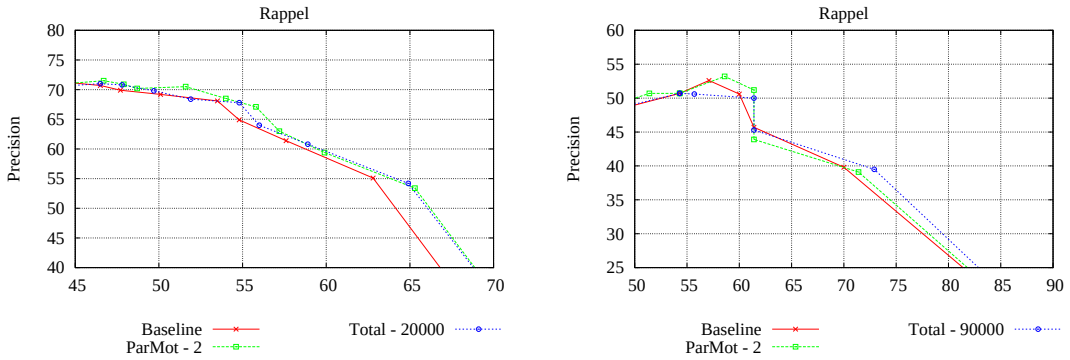
Si les mesures de confiance peuvent être également prises en compte lors du calcul de la probabilité généralisée, ce n'est pas le cas des relations sémantiques. En effet, dans ce cas, l'intégration de relations sémantique reviendrait à multiplier la probabilité d'apparition d'un mot u dans un segment S_i par la somme des relations que ce mot entretient avec les autres mots de S_i, ce qui n'a pas de sens.

Il est important de noter que, grâce à notre technique d'intégration, si les relations sémantiques utilisées ne sont pas adaptées à un document particulier – par exemple, lors de la segmentation d'un document traitant d'un domaine n'apparaissant pas dans le corpus employé lors de l'apprentissage des relations sémantiques – la valeur de $C_i''(u)$ ne sera pas différente de celle de $C_i(u)$, $r(u,v)$ étant nul pour tout couple de mots u et v ne partageant pas de lien sémantique. En d'autres termes, des relations sémantiques hors-domaine n'auront pas d'impact sur la segmentation thématique, contrairement à ce que l'on peut observer avec des approches d'analyse sémantique latente (Deerwester et al., 1990; Landauer et al., 1998).

Les relations sémantiques utilisées, dont l'extraction est décrite dans le chapitre 3, peuvent être sélectionnées grâce à deux techniques différentes $Total_{\rho_1}$, qui consiste à conserver les ρ_1 meilleures relations, tous mots confondus, et $ParMot_{\rho_2}$, qui sélectionne les ρ_2 meilleures relations pour chaque mot, éventuellement associées à une méthode de filtrage. Les nombreux tests effectués, concernant le type et le nombre de relations introduites, les méthodes de sélection employées, *etc.*, ont produit une quantité importante de résultats que nous ne présentons pas *in extenso* dans ce chapitre. Nous décrivons ici seulement l'influence générale de l'intégration de relations sémantiques dans l'algorithme de segmentation, plus de détail étant proposés dans l'annexe B.

Comme nous l'avons souligné précédemment, l'intégration de relations sémantiques dans l'algorithme de segmentation vise un double objectif. Premièrement ces relations doivent permettre de gérer la faible répétition de vocabulaire présent dans nos données télévisuelles, et notamment dans notre corpus de journaux télévisés. Deuxièmement, nous souhaitons montrer que l'utilisation de relations sémantiques pénalise les mots mal transcrits dans le calcul de la cohésion lexicale et atténue ainsi l'impact des erreurs de transcription présentes dans nos corpora.

Afin de tester l'influence des relations sémantiques sur la faible répétition de vocabulaire, nous avons comparé leur impact sur les résultats d'une segmentation thématique appliquée sur les deux types d'émissions. La figure 5.4 présente les meilleurs résultats obtenus lors de l'intégration de relations paradigmatiques sélectionnées grâce aux techniques *ParMot* et *Total*. Sur ces courbes rappel/précision, nous remarquons que les relations sémantiques ont moins d'influence sur le corpus d'émissions de reportages que sur les journaux télévisés. Cette observation nous laisse penser que les relations sémantiques introduites permettent de pallier une faible répétition du vocabulaire, ce qui explique qu'elles ont une influence moindre sur le corpus d'émissions de reportages, associé à une répétition de vocabulaire cinq fois plus élevée que celle présente dans le corpus de *JT*. Cette différence peut également être liée au corpus d'apprentissage à partir duquel ont été extraites les relations. Ce corpus est en effet composé d'articles de journaux et se rapproche donc davantage du corpus de journaux télévisés que de celui d'émissions de reportages. Cependant, il est important de noter que si les relations sémantiques ont moins d'influence sur le corpus *Sept à Huit*, elles ne dégradent pas pour autant la qualité de la segmentation.



corpus *JT* corpus *Sept à Huit*

FIG. 5.4 – Prise en compte des relations sémantiques lors de la segmentation thématique de journaux télévisés et d'émissions de reportages *Sept à Huit*

TAB. 5.1 – Influence des relations sémantiques sur les erreurs de transcription

	Sans relations	$ParMot_2$
Transcriptions automatiques[3]	65,56	68,94
Transcriptions manuelles	72,96	73,3

Le second objectif des relations sémantiques consiste à pallier les erreurs de transcriptions en diminuant le poids des mots mal transcrits dans le calcul de la cohésion lexicale. Nous pensons, en effet, que les mots erronés ont peu de chances d'être reliés sémantiquement aux autres mots du segment et ne seront ainsi pas mis en avant par les relations sémantiques, contrairement aux mots correctement transcrits. Afin de tester cette hypothèse, nous avons pris en compte des relations paradigmatiques, sélectionnées grâce à la méthode *ParMot*, lors de la segmentation de journaux télévisés transcrits manuellement. Dans ce cas, l'intégration des relations sémantiques ne permet d'augmenter la valeur de la mesure F_1 que de 0,34 point seulement (*cf.* tableau 5.1). Si l'on compare ce gain avec celui observé sur les émissions transcrites automatiquement, nous constatons qu'il est presque 10 fois plus élevé pour les transcriptions automatiques, ce qui nous permet d'affirmer que l'intégration de relations sémantiques dans l'algorithme de segmentation thématique est effectivement utile pour pallier les erreurs de transcription présentes dans nos données.

5.1.3 Interpolation

Dans les émissions télévisuelles, et particulièrement dans les journaux télévisés, les segments thématiques peuvent être très courts. Une critique qui peut être formulée à l'égard de la méthode de calcul de la cohésion lexicale développée dans (Utiyama and Isahara, 2001) est que, pour ces petits segments, le modèle de langue Δ_i risque d'être mal estimé. Afin de gérer ces petits segments, la technique originale que nous proposons consiste à utiliser une estimation plus sophistiquée des modèles de langue, obtenue en interpolant les modèles de langue appris au niveau des segments avec un modèle de langue estimé sur la transcription entière.

Deux techniques d'interpolation ont été employées – l'interpolation de probabilité (Jelinek and Mercer, 1981) et l'interpolation des comptes des mots (Bacchiani and Roark, 2003).

La première méthode d'interpolation testée consiste à interpoler les probabilités. Dans ce cas, la cohésion lexicale d'un segment S_i est mesurée de la façon suivante :

$$\begin{aligned} \ln P[S_i; S_i, T] &= \sum_{j=1}^{n_i} \ln(\lambda P[w_j^i; \Delta_i] + (1-\lambda) P[w_j^i; \Delta_t]) \\ &= \sum_{j=1}^{n_i} \ln \left(\lambda \frac{C_i(w_j^i) + \xi}{\sum_{u \in V_T} C_i(u) + \xi} + (1-\lambda) \frac{C_t(w_j^i)}{\sum_{u \in V_T} C_t(u)} \right) , \end{aligned} \tag{5.4}$$

avec Δ_i le modèle de langue estimé sur le segment S_i et Δ_t celui calculé sur la transcription de l'émission complète T. $C_t(u)$ est le compte d'un mot u dans T et $C_i(u)$ le compte de ce mot dans S_i. ξ est un lissage correspondant au lissage de Laplace lorsque $\xi = 1$.

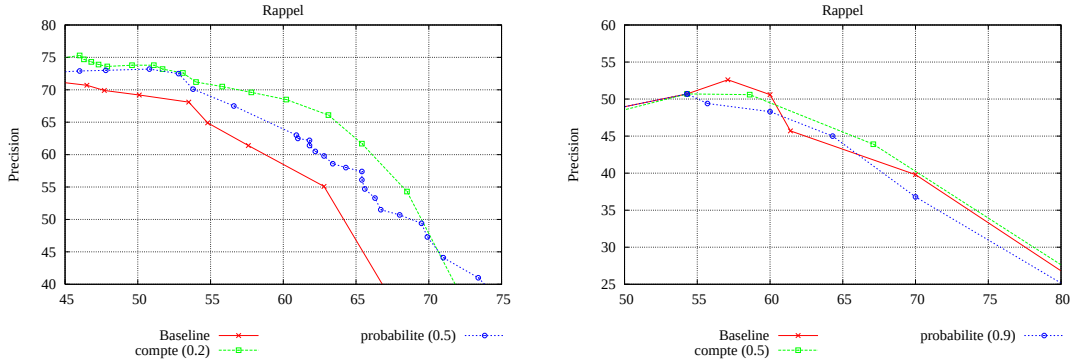
Plutôt que d'interpoler les probabilités, l'interpolation des modèles de langue peut également être effectuée à partir de l'interpolation des comptes des mots. Dans ce cas, la cohésion lexicale d'un segment S_i est définie par :

$$\begin{aligned} \ln P[S_i; S_i, T] &= \sum_{j=1}^{n_i} \ln P[w_j^i; \Delta_{it}] \\ &= \sum_{j=1}^{n_i} \ln \left(\frac{\lambda(C_s(w_j^i) + \xi) + (1-\lambda) C_t(w_j^i)}{\sum_{u \in V_T} \lambda(C_i(u) + \xi) + (1-\lambda) C_t(u)} \right) , \end{aligned} \tag{5.5}$$

avec Δ_{it} le modèle de langue du segment S_i interpolé avec celui de la transcription T. Il est à noter que, comme pour l'interpolation des probabilités, les mots apparaissant fréquemment dans T vont être associés à une forte probabilité, indépendamment de leur fréquence dans S_i, alors que ceux peu fréquents dans T seront toujours associés à une faible probabilité, liée à λ. Cependant, grâce à la renormalisation par la somme de tous les comptes, cette observation sera probablement moins dommageable dans le cas de l'interpolation des comptes. De ce fait, cette technique est, selon nous, plus susceptible de fournir des résultats cohérents avec ceux attendus.

Les résultats obtenus grâce aux deux méthodes d'interpolation des modèles de langue sont présentés sur la figure 5.5. La courbe rouge représente la segmentation calculée avec une estimation classique des modèles de langue tandis que les deux autres courbes correspondent aux différentes techniques d'interpolation. Nous constatons que, pour les deux corpora, l'interpolation des comptes fournit les meilleurs résultats. En effet, pour le corpus de journaux télévisés, la valeur de la mesure F_1 est augmentée de 4,9 points lors de l'interpolation des comptes contre 2,3 points pour l'intégration des probabilités, augmentations statistiquement significatives selon le test de Student. Ces valeurs sont, par ailleurs, grandement diminuées

[3]La valeur de la mesure F_1 associée aux transcriptions automatiques est différente de celle affichée dans les tableaux de l'annexe B. Cette différence s'explique par le fait que les expériences reportées dans ce tableau ont été effectuées sur un corpus composé des 8 journaux télévisés correspondant à ceux transcrits manuellement.



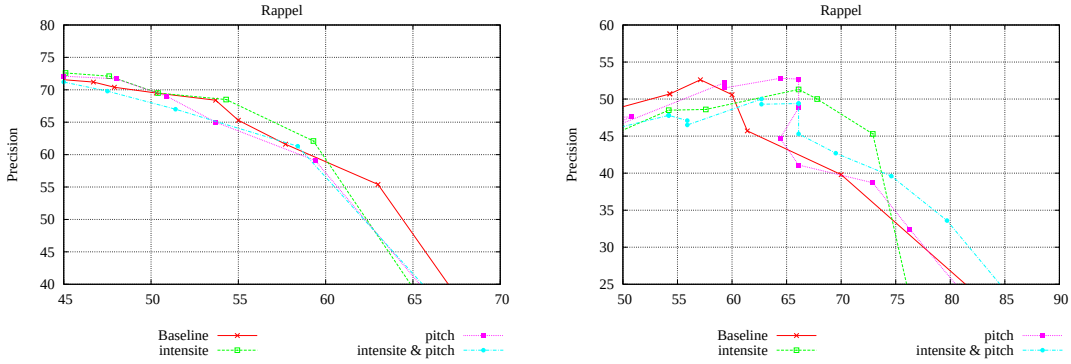
corpus *JT* corpus *Sept à Huit*

FIG. 5.5 – Interpolation des modèles de langue pour améliorer la segmentation thématique de journaux télévisés et d'émissions de reportages *Sept à Huit*. Pour les deux techniques d'interpolation des modèles de langue, la valeur entre parenthèses correspond à la valeur optimale du paramètre ξ

pour le corpus *Sept à Huit* puisque l'interpolation des comptes améliore la qualité de la segmentation de 0,7 point seulement, tandis que l'interpolation des probabilités n'offre pas de gain comparativement à la segmentation classique. Ce comportement des deux techniques d'interpolation des modèles de langue s'explique par le fait que, comme nous l'avons mentionné plus tôt, l'interpolation des modèles de langue associe aux mots fréquents du texte des valeurs de probabilité fortes quelle que soit leur fréquence dans le segment, alors que les mots peu fréquents sont toujours associés à une valeur de probabilité faible. Ce comportement n'est cependant pas aussi marqué dans le cadre de l'interpolation des comptes, grâce à la normalisation proposée dans l'équation (5.5), ce qui rend cette méthode plus efficace. Les courbes rappel/précision, présentées en figure 5.5 montrent également que l'interpolation des modèles de langue a beaucoup moins d'impact sur le corpus d'émissions de reportages que sur celui composé de journaux télévisés. Cette remarque s'explique aisément par le fait que la longueur moyenne des segments dans les émissions de reportages est beaucoup plus longue (8,6 minutes) que celle des segments des journaux télévisés (1,6 minute). De ce fait, les modèles de langue calculés pour représenter les segments dans la méthode de segmentation de base sont déjà bien estimés dans le cas des émissions *Sept à Huit*.

5.2 Utilisation de la prosodie

Dans cette section, nous souhaitons tirer parti de la modalité audio de nos données télévisuelles. Le but de l'intégration d'informations prosodiques dans le calcul de la cohésion lexicale consiste à traduire l'intention du locuteur et à augmenter l'importance des mots proéminents dans le discours, généralement associés à une valeur informative importante. Favoriser ces mots proéminents a pour objectif de calculer des modèles de langue plus représentatifs des segments et d'obtenir ainsi une évaluation plus juste de la cohésion lexicale. S'il a été montré dans (Tür et al., 2001) que la combinaison d'indices textuels et prosodiques permet d'améliorer la qualité de la segmentation thématique de documents en langue anglaise – comparativement à la segmentation opérée avec les deux indices pris séparément –, aucune étude n'a été effectuée, à notre connaissance, sur un corpus français.



corpus *JT* corpus *Sept à Huit*

FIG. 5.6 – Prise en compte d'informations prosodiques lors de la segmentation thématique de journaux télévisés et d'émissions de reportages *Sept à Huit*

L'intégration des informations prosodiques se fait de la même manière que la prise en compte des mesures de confiance. Durant la phase d'estimation des modèles de langue, le compte des mots $C_i(u)$ est remplacé par la somme des valeurs de scores acoustiques associées à chacune des occurrences de ce mot.

$$C_i'''(u) = \sum_{w_j^i = u} a(w_j^i) \ , \tag{5.6}$$

avec $a(w_j^i)$ le score acoustique de w_j^i calculé grâce à la technique d'extraction d'informations acoustiques décrite en section 3.2. De cette manière, les mots faiblement proéminents auront moins d'impact dans le calcul de la cohésion lexicale.

Lorsque les informations prosodiques sont intégrées lors du calcul de la probabilité généralisée, la log-probabilité de l'occurrence d'un mot est multipliée par la valeur du score acoustique de l'occurrence de ce mot :

$$\ln P[S_i; \Delta_i] = \sum_{j=1}^{n_i} a(w_j^i) \ln P[W_j^i; \Delta_i] \ . \tag{5.7}$$

Finalement, ces informations acoustiques peuvent également être prises en compte à la fois durant l'estimation du modèle de langue et durant le calcul de la probabilité généralisée en remplaçant $C_i(u)$ par $C_i'''(u)$ dans l'estimation du modèle de langue.

Les informations acoustiques employées ayant été extraite pour chaque 0,01 seconde du signal, et la durée de prononciation d'un mot étant généralement supérieure à cette valeur, quatre stratégies d'alignement entre ces valeurs et les transcriptions ont été utilisées : MAX qui associe à chaque mot la valeur maximale parmi toutes les informations acoustiques observée pour la durée du mot, MOYENNE et ET qui calculent la moyenne et l'écart-type et MIN qui conserve la valeur minimum. Ces techniques d'alignement, combinées aux différents types d'informations acoustiques, intensité ou pitch, ont conduit à un nombre important d'expérimentations. Comme pour les relations sémantiques, nous présentons dans ce chapitre les principales tendances observées lors de l'intégration des informations prosodiques, plus de détail pouvant être trouvés dans l'annexe B.

Les courbes rappel/précision, présentées sur la figure 5.6, montrent tout d'abord que l'intégration d'informations prosodiques a un impact différent sur la qualité de la segmentation en

fonction du corpus considéré. En effet, si la prosodie permet d'améliorer de façon statistiquement significative les performances de l'algorithme de segmentation dans le cas des émissions de reportages – la valeur de la mesure F_1 est augmentée de 3,7 points lorsque des valeurs de *pitch* sont prises en compte lors du calcul du modèle de langue – seul un gain de 1 point est constaté pour le corpus de journaux télévisés. L'écart de performance entre les deux corpora peut être lié, d'une part, à la différence existant dans l'accentuation de la parole au sein des deux émissions, les journalistes commentant les émissions de reportages adoptant un ton très différent de celui employé dans les journaux télévisés. D'autre part, si l'amélioration observée pour les émissions *Sept à Huit* peut être due à l'utilisation de mots plus caractéristiques lors de l'évaluation de la cohésion lexicale, nous soupçonnons également, au vu des résultats obtenus sur les journaux télévisés, que les informations prosodiques agissent de la même façon que les mesures de confiance et que la différence constatée entre les deux corpora s'explique par l'écart entre leur taux d'erreur. Nous pensons en effet qu'il existe une forte corrélation entre les valeurs des mesures de confiance et celles des informations prosodiques, les mots mal reconnus par le système de transcription étant associés à une valeur prosodique faible, tandis que ceux dont la prononciation a été accentuée par le locuteur ont plus de chance d'être mieux reconnu.

5.3 Bilan du chapitre

Dans ce chapitre, nous avons proposé différentes techniques permettant de rendre le critère de la cohésion lexicale plus robuste aux spécificités de programmes télévisuels transcrits, tout en restant suffisamment générique pour être applicable à différents types de données. Nous avons ainsi montré que les relations sémantiques et l'interpolation des modèles de langue améliorent la qualité de la segmentation thématique d'émissions composées de petits segments thématiques et contenant une faible répétition de vocabulaire. De plus, une comparaison entre l'impact des relations sémantiques sur les transcriptions manuelles et automatiques a permis de démontrer que leur intégration atténue l'influence des erreurs de transcription sur le calcul de la cohésion lexicale. Les expériences menées ont également mis en évidence le fait que les mesures de confiance rendent la cohésion lexicale moins sensible à la présence de mots erronés, ces mesures de confiance étant par ailleurs plus efficaces pour des transcriptions associées à un taux d'erreur élevé. L'utilisation d'informations prosodiques sur un corpus francophone a finalement montré que la prise en compte de la multimodalité des données augmente les performances de l'algorithme de segmentation, de façon plus ou moins importante selon le type de programme télévisé considéré.

Si ces différents éléments peuvent être utilisés séparément, il est également possible de les combiner – les différentes combinaison des mesures de confiance, des relations sémantiques et des techniques d'interpolation des modèles de langue sont présentées en détail dans l'article (Guinaudeau et al., 2011) – mais également de les associer à des informations de rupture de la cohésion lexicale telles que présentées dans le chapitre précédent. De légères améliorations ont ainsi pu être observées lors de la combinaison de certains de ces indices. Par exemple, l'utilisation conjointe de l'interpolation et des mesures de confiance améliore la qualité de la segmentation des journaux télévisés par rapport à leur utilisation séparée. De même, la combinaison des informations de rupture, des mesures de confiance et de l'interpolation améliore significativement la précision obtenue, pour une même valeur de rappel (45,1), par rapport à l'utilisation indépendante de ces trois indices. La précision est augmentée de 0,4

point par rapport à l'ajout d'informations de rupture, de 9 points par rapport à l'utilisation des mesures de confiance et de 6,5 points comparativement à l'interpolation des modèles de langue. Certaines combinaisons n'apportent cependant aucun gain. C'est le cas, par exemple, de la prise en compte conjointe des relations sémantiques et de l'interpolation, qui semble redondante, les deux indices visant le même objectif, c'est-à-dire la gestion de courts segments thématiques. Dans les cas où aucune amélioration n'est observée, il est cependant important de noter qu'aucune dégradation de la qualité de la segmentation thématique n'a été constatée.

Les techniques développées pour adapter le critère de la cohésion lexicale aux spécificités des données télévisuelles produisant des résultats encourageants, plusieurs pistes peuvent être envisagées pour les consolider. Premièrement, l'utilisation d'informations prosodiques offrant un gain de performance important pour le corpus d'émissions de reportages, il semble nécessaire d'éclaircir les raisons de cet apport en étudiant les résultats d'une segmentation thématique prenant en compte la prosodie sur un corpus transcrit manuellement. Les transcriptions manuelles utilisées jusqu'à présent ne possédant pas d'informations temporelles associées à chacun des mots, aucune information prosodique n'a pu leur être attachée. Or, afin de différencier le rôle de la prosodie de celui des mesures de confiance, il nous semble important d'étudier leur éventuel apport sur des transcriptions ne contenant pas d'erreurs de transcription. Deuxièmement, la qualité et le nombre des relations sémantiques introduites ayant un impact important sur la qualité de la segmentation produite, il pourrait s'avérer judicieux d'acquérir des relations sémantiques par le biais de méthodes d'extraction plus sophistiquées et d'évaluer l'impact des méthodes utilisées sur les performances de l'algorithme de segmentation. Finalement, l'utilisation des mesures de confiance ayant une influence positive sur la qualité de la segmentation thématique, nous pensons qu'il pourrait être intéressant d'appliquer l'algorithme de segmentation thématique sur les sorties intermédiaires proposées par le système de reconnaissance automatique de la parole, comme les graphes de mots ou les réseaux de confusion. (Mohri et al., 2010) a en effet montré que l'utilisation des hypothèses de transcription plutôt que la transcription finale augmentait les performances d'un algorithme de segmentation thématique fondé sur la détection de ruptures.

L'adaptation de la mesure de la cohésion lexicale aux particularités de nos données télévisuelles fournissant de bons résultats, nous proposons dans la suite de cette thèse d'utiliser cette phase de segmentation thématique comme première étape de structuration. La troisième et dernière partie de ce manuscrit est consacrée à la présentation de deux méthodes de structuration automatique de programmes télévisés. Dans le chapitre 6, nous décrivons les techniques mises en place pour produire une *structuration thématique linéaire* visant à mettre en relation des segments thématiquement homogènes extraits d'une collection de documents. Le chapitre 7 est, quant à lui, consacré à un travail plus exploratoire sur la *structuration thématique hiérarchique*, les programmes télévisés sur lesquels nous avons appliqué l'algorithme de segmentation thématique linéaire possédant une structure hiérarchique importante qui n'a, jusqu'ici, pas été prise en compte.

Troisième partie

Structuration d'émissions

Chapitre 6

Mise en relation de segments thématiquement homogènes

Ce chapitre poursuit l'étude des différentes étapes de structuration automatique de documents audiovisuels en abordant la tâche de *structuration thématique linéaire*. Cette tâche de structuration a pour objectif de mettre en relation des éléments extraits d'une collection de documents qui abordent des thématiques similaires. Dans notre contexte, elle consiste à associer deux reportages tirés de journaux télévisés ou d'émissions de reportages, dès lors qu'ils traitent d'un fait d'actualité de même nature. La mise en place de cette méthode de *structuration thématique linéaire* doit, par exemple, permettre à des utilisateurs d'accéder à tous les éléments d'une collection de documents qui traitent d'un sujet qui les intéresse ou de suivre les évolutions d'un fait d'actualité.

Afin de mettre en place une telle structuration, nous proposons, dans ce chapitre, une méthode associant des segments, obtenus à partir d'une phase de segmentation thématique préalable, partageant une même thématique. Cette association repose principalement sur la comparaison des mots clés caractéristiques des segments, extraits des transcriptions automatiques de la parole qu'ils contiennent. Dans la section 6.1, nous exposons, après la description des différentes méthodes disponibles dans l'état de l'art, la technique utilisée pour comparer nos segments thématiques. Cependant, si les transcriptions automatiques employées dans notre approche permettent d'accéder au contenu sémantique des reportages considérés, elles n'autorisent pas la prise en compte des spécificités des programmes télévisés. Afin d'adapter notre méthode aux particularités de nos données audiovisuelles, nous proposons, dans la section 6.2 de ce chapitre, deux modifications.

Premièrement, nous pensons que l'utilisation d'indices propres à l'oral doit permettre d'améliorer la mise en relation de segments thématiquement homogènes dans le cadre d'émission de télévision. La méthode de base reposant sur la comparaison des mots clés caractéristiques, nous souhaitons prendre en compte les mots accentués de façon intentionnelle par le locuteur – probablement associés à une information sémantique forte – lors de la caractérisation des segments. Pour ce faire, nous utilisons des informations prosodiques, extraites automatiquement de nos données, pour modifier la représentation de nos segments thématiques. Cette modification, ainsi que les résultats obtenus grâce à elle sur nos deux corpora de test, sont présentés dans la section 6.2.1.

Deuxièmement, le calcul de la similarité entre deux segments thématiques employé dans la méthode de base peut être amélioré en y intégrant des relations sémantiques. L'objectif de

ces relations est, d'une part, de faire le lien entre deux segments abordant le même sujet par le biais d'un vocabulaire différent. D'autre part, la prise en compte des liens sémantiques existant entre les différents mots caractéristiques des segments doit permettre de pallier les erreurs de transcription présentes dans nos données, comme nous le présentons dans la section 6.2.2.

Finalement, nous décrivons dans ce chapitre deux applications développées grâce à la technique de segmentation thématique décrite dans les chapitres précédents et à la mise en relations de segments thématiques. Ces applications, toutes deux employées sur des journaux télévisés, sont détaillées dans la section 6.3.

6.1 Structuration par la mise en relations de segments thématiques : principe

6.1.1 État de l'art

Le suivi de sujet ou d'événement au sein d'une collection de documents a largement été étudié à travers le projet de recherche *Topic Detection and Tracking* (TDT) lancé en 1997. L'objectif de ce projet consiste à développer des méthodes permettant de mettre en relation des segments de document textuels ou audiovisuels (par le biais de transcriptions) abordant des sujets similaires. Cette mise en relation doit permettre aux utilisateurs de suivre les évolutions d'un événement au cours du temps ou de visualiser les différences de traitement d'un même fait d'actualité par différents médias. Les approches proposées dans ce cadre reposent généralement sur une segmentation thématique des documents d'une part, qui permet d'extraire des parties de documents homogènes du point de vue du sujet abordé, et sur le calcul de la similarité entre ces segments d'autre part. Les méthodes de calcul de la similarité entre documents ou segments de documents se basent généralement sur la proportion de vocabulaire partagé par deux segments. Plus les segments ont de termes en commun, plus ils ont une probabilité élevée d'aborder le même sujet. Deux types d'approches sont privilégiées pour évaluer la proximité de vocabulaire entre deux segments : les approches vectorielles (Ide et al., 2004) et les approches à base de modèles de langue statistiques (Mulbregt et al., 1998).

Les approches vectorielles utilisent des méthodes classiquement mises en œuvre en recherche d'information (RI) pour apparier une requête aux documents de la collection et ainsi retourner une liste de résultats pertinents. Pour cela, les documents et requêtes sont représentés par des vecteurs de termes pondérés, appelés termes d'indexation, et la proximité sémantique entre vecteurs est calculée grâce à une mesure de similarité. Dans le cadre du suivi d'événement, ces vecteurs peuvent être composés de noms de personnes (Ide et al., 2005) ou d'entités nommées (Ide et al., 2004), pondérés par leur fréquence. Cependant, la grande majorité des études se base sur des vecteurs de mots pondérés par le critère *tf-idf* (Yang et al., 1999), éventuellement améliorés grâce à des méthodes de traitement automatique des langues. Certaines de ces techniques sont devenues des références, telles que l'utilisation des racineurs – permettant de mettre en relation des mots graphiquement différents mais sémantiquement proches – ou de lemmatiseurs. Des relations sémantiques, obtenues manuellement ou acquises automatiquement, sont également largement employées pour mettre en relation des (parties de) documents abordant des sujets similaires sans partager de vocabulaire commun. Finalement, les méthodes de recherche d'information utilisent également des techniques plus évoluée que le simple *bag-of-word* pour caractériser les documents et intègrent dans les termes d'indexation une notion d'ordre entre les mots, que ce soit par le biais d'extraction de termes ou de structures complexes. Cependant les erreurs de transcription ou la nature souvent moins

structurée de la parole prononcée rendent ces dernières techniques souvent peu adaptées au traitement des documents oraux.

Pour s'adapter aux transcriptions automatiques de la parole contenue dans les documents audiovisuels, de nombreux travaux de suivi d'événement se sont inspirés de techniques mises en place dans le cadre des campagnes de recherche d'information dans les documents oraux (*Spoken Document Retrieval*) durant lesquelles l'indexation de documents audios à partir des transcriptions automatiques de la parole a été largement étudiée (Garofolo et al., 2000). La plupart des techniques de *Spoken Document Retrieval* cherchent à pallier les particularités des transcriptions automatiques en adaptant les techniques de recherche d'information classique à ces spécificités. Afin de gérer le problème des mots hors-vocabulaire, c'est-à-dire des mots prononcés dans l'émission qui n'appartiennent pas au lexique du système de reconnaissance automatique de la parole, certains travaux mettent en place des techniques d'extension de documents. Dans (Singhal et al., 1999), cette extension de documents est effectuée grâce à des mots extraits de documents proches sémantiquement et qui n'apparaissent pas dans le document considéré. Les auteurs proposent également d'utiliser les réseaux de confusion obtenus grâce au système de transcription afin de prendre en compte les mots similaires d'un point de vue phonétique qui auraient pu être prononcés à la place du mot qui apparaît dans l'hypothèse de transcription. Suivant une philosophie similaire, Johnson *et al.* (Johnson et al., 2000) ont montré qu'un taux d'erreur pouvant aller jusqu'à 20% peut être compensé grâce à une extension de documents utilisant des relations sémantiques extraites de WordNet (Miller, 1995) ou des relations existant entre localisations géographiques définies manuellement.

Dans le cadre de suivi d'événements dans des documents audiovisuels, certains travaux tirent parti de la multimodalité des données en intégrant dans leur système des indices vidéos. Dans (Hsu and Chang, 2006) par exemple, les auteurs détectent des concepts visuels définis dans le cadre de la campagne d'évaluation TRECVID pour mettre en relation des segments abordant des sujets similaires. Cependant, ils montrent que l'apport de l'indice visuel est très faible comparé aux résultats obtenus avec la modalité textuelle.

Finalement, dans (Allan, 2002b), Allan propose plusieurs pistes visant à améliorer l'indexation de documents oraux qui peuvent être utilisées dans le cadre de la mise en relation de documents audiovisuels. Il suggère, par exemple, d'intégrer les mesures de confiance fournies par le système de reconnaissance automatique de la parole dans le calcul des pondérations associées aux termes d'indexation d'un document transcrit. Il propose également d'employer des informations prosodiques pour prendre en compte la façon dont les mots sont prononcés dans le document, la prosodie pouvant traduire l'importance d'un terme vis-à-vis du thème abordé. Il a en effet été montré (Crestani, 2001) qu'il existe, dans la langue anglaise, un lien direct entre l'accentuation acoustique d'un terme et sa valeur informative. De plus, les auteurs de (Chen et al., 2001) concluent dans leurs travaux que l'utilisation d'informations prosodiques apporte une légère amélioration dans un système de recherche d'information dans des documents oraux développé pour la langue chinoise.

6.1.2 Méthode retenue

Pour atteindre notre objectif de structuration thématique linéaire, nous avons choisi de mettre en place une technique classiquement utilisée en recherche d'information. Notre système de mise en relation des segments thématiquement homogènes, similaire à celui présenté dans (Yang et al., 1999), se divise en deux étapes (*cf.* Figure 6.1). Premièrement, chaque segment est représenté par un vecteur de mots, pondérés par leur score *tf-idf*, puis une mesure

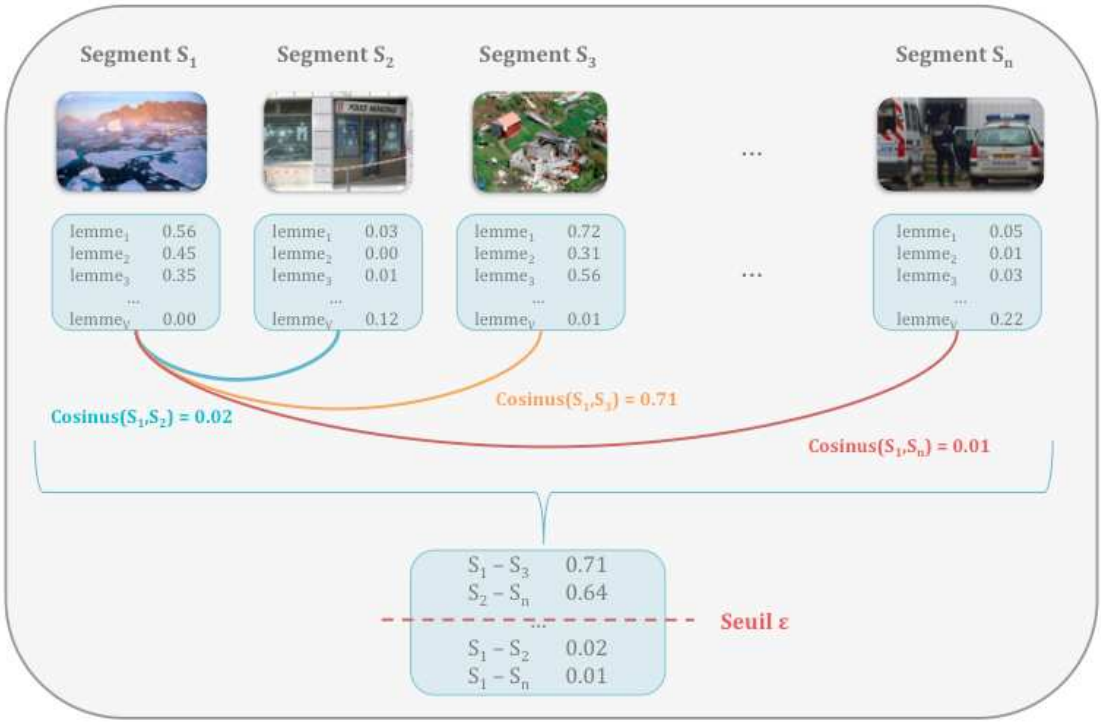


FIG. 6.1 – Mise en relation de segments thématiquement homogènes

de similarité est calculée entre chaque couple possible de vecteurs afin d'évaluer la proximité thématique de chaque couple de segments. Les paires de segments dont les vecteurs sont associés à une mesure de similarité supérieure à un seuil ϵ sont finalement considérés comme liés sémantiquement par notre système.

L'évaluation de la similarité entre segments thématiques reposant essentiellement sur la comparaison de leurs vecteurs représentatifs, la qualité de ces vecteurs est un élément crucial du système. Le calcul des vecteurs caractéristiques consiste à associer à chaque mot du corpus un score traduisant son importance au sein du segment. Nous avons choisi dans ce système d'utiliser le critère *tf-idf*, couramment utilisé en recherche d'information pour traduire la capacité d'un mot à discriminer le document dans lequel il se trouve par rapport à une collection de textes de référence. Étant donné un mot w d'un document d, la pondération *tf-idf* est calculée comme le produit d'une pondération locale $tf(w,d)$ qui exprime la fréquence de w dans d, et d'une pondération globale $idf(w,\Omega)$ liée au nombre d'occurrences de w dans les documents de la collection de référence[1]. Cette pondération globale mesure l'importance du terme w dans l'ensemble de la collection et permet de favoriser les mots apparaissant fréquemment dans le document mais peu dans la collection de référence, considérés comme plus discriminants.

Le score *tf-idf* a été calculé grâce à l'outil *kiwi* développé par Lecorvé *et al.* (Lecorvé et al., 2008). Parmi les différentes variantes existant pour la mise en œuvre des pondérations

[1]La collection de référence utilisée dans ce travail est composée de 800 000 articles extraits du journal *Le Monde* entre 1987 et 2003. Ce corpus ayant servi à estimer les probabilités du modèle de langue et le vocabulaire du système de transcription automatique de la parole employé pour transcrire nos données, le vocabulaire apparaissant dans nos transcriptions n'est pas très différent de celui de la collection de référence.

locales et globales (Singhal, 1997), nous utilisons dans ces travaux celles définies de la façon suivante :

$$tf(w,d) = \frac{freq(w,d)}{\max\limits_{x \in d} freq(x,d)} \quad \text{avec} \quad freq(w,d) = \frac{|\, w \,|_d}{|\, d \,|} \tag{6.1}$$

et

$$idf(w,\Omega) = \log \frac{|\, \Omega \,|}{|\, w \,|_\Omega} \tag{6.2}$$

où $|\, d \,|$ désigne le nombre de mots de d, $|\, \Omega \,|$ désigne le nombre de documents dans la collection, et $|\, w \,|_d$ et $|\, w \,|_\Omega$ désignent respectivement le nombre d'occurrences de w dans d et le nombre de documents de Ω qui contiennent le mot w. Finalement, une dernière étape de normalisation nous permet d'obtenir le score inspiré de la recherche d'information, $S_{ir}(w)$, compris entre 0 et 1 :

$$S_{ir}(w) = \frac{tf(w,d) \times idf(w,\Omega)}{\max\limits_{x \in d}(tf(x,d) \times idf(x,\Omega))} \tag{6.3}$$

Pour chacun de nos corpora, ce score est calculé pour chaque mot[2] du vocabulaire, c'est-à-dire l'ensemble des mots apparaissant au moins une fois dans les segments. Chaque segment étant ainsi caractérisé par un vecteur de mots associés à une pondération *tf-idf*. De plus, certains segments thématiques pouvant être très courts et ne contenir que peu de répétitions de mots, nous appliquons en aval une lemmatisation sur chacune de nos transcriptions afin de rassembler les mots partageant un même lemme.

Finalement, la proximité thématique des segments est évaluée grâce à la mesure cosinus. Cette mesure de similarité renvoie, pour les deux vecteurs représentatifs des segments A et B, une valeur comprise entre 0 et 1, 1 correspondant à une similarité très forte entre les deux segments. Formellement, cette mesure est définie par :

$$cos(A,B) = \frac{\sum_{j=1}^{L} a_j b_j}{\sqrt{\sum_{j=1}^{L} a_j^2} \sqrt{\sum_{j=1}^{L} b_j^2}} \ , \tag{6.4}$$

avec L le nombre de lemmes sélectionnés, associés aux scores les plus élevés, pour caractériser le contenu des segments. Le calcul de la similarité entre les différents segments a été effectué à l'aide de vecteurs composés de 100 lemmes. Diverses expériences ont, en effet, montré que les résultats obtenus étaient meilleurs avec cette valeur, bien que la différence entre les performances ne soit plus significative dès lors que la valeur de L est supérieure à 60, et ceci pour les deux corpora.

Ce système de mise en relation de segments thématiques a été testé sur des segments extraits de 8 journaux télévisés et de 13 émissions de reportages *Sept à Huit*. Ces segments correspondant aux reportages éventuellement associés aux plateaux de lancement et de fin, tel que présenté dans la section 4.1.2, auraient pu être extraits de façon automatique grâce aux techniques décrites dans les deux chapitres précédents. Cependant, nous avons choisi d'opérer cette extraction de façon manuelle afin d'exclure de notre analyse des résultats les effets induits par de potentielles erreurs de segmentation. Nous obtenons grâce à cette segmentation manuelle 177 segments pour les journaux télévisés et 72 pour les émissions *Sept à Huit*. La mise en relation de référence a été effectuée manuellement en reliant deux segments abordant des sujets proches, deux segments traitant de la campagne présidentielle étant, par exemple, considérés comme reliés sémantiquement, contrairement à deux segments discutant

[2] À l'exception des mots vides et des verbes modaux tels que *faire*, *avoir*, *etc.*

de politique de manière plus générale. 205 liens ont ainsi été créés pour le corpus de journaux télévisés et 56 pour les émissions de reportages. Les résultats obtenus par le système sont évalués grâce à des mesures de rappel, représentant la proportion de liens sémantiques pertinents retrouvés, et de précision, traduisant la proportion de liens sémantiques pertinents parmi ceux retrouvés.

6.2 Mise en relation de segments de programmes TV

Le critère *tf-idf* ayant été développé pour la recherche d'information au sein de documents textuels, il ne permet pas de considérer les particularités des transcriptions de programmes télévisés, telles que les erreurs de transcriptions et le faible taux de répétition dans nos segments, ainsi que l'aspect oral de nos données. Pour tenir compte de ces spécificités, deux modifications du système précédemment décrit ont été envisagées : la modification de la représentation vectorielle des segments et l'intégration de relations sémantiques dans le calcul de la similarité entre les vecteurs.

6.2.1 Modification de la représentation vectorielle

Si le critère *tf-idf* utilisé pour la caractérisation des segments thématiques prend en compte la fréquence des mots apparaissant dans les transcriptions, il ne permet pas de traduire l'importance donnée à un mot par un locuteur lors de son énonciation.

Dans cette section, nous étudions l'impact de l'utilisation d'informations prosodiques dans le calcul des vecteurs caractéristiques des segments thématiques. Nous croyons que les mots les plus caractéristiques d'un segment sont prononcés avec plus d'emphase ou d'intensité que les mots moins importants. En effet, de nombreux travaux ont montré que la prosodie d'un mot est corrélée avec sa valeur informative, le fait qu'un mot soit prononcé avec plus d'emphase par un locuteur pouvant traduire le fait que ce mot revêt une importance particulière à ses yeux.

Pour explorer cette hypothèse, nous avons tout d'abord associé des scores acoustiques à chacun des lemmes apparaissant dans les transcriptions de nos émissions télévisuelles. Ces scores, calculés grâce à la technique décrite en 3.2, sont compris entre 0, pour les mots faiblement proéminents dans le discours, et 1, pour ceux prononcés avec le plus d'emphase. Chaque lemme pouvant se répéter plusieurs fois au sein d'une transcription, nous avons mis en place deux stratégies pour gérer la répétition de vocabulaire. Dans la première stratégie, *-M-*, la valeur maximale parmi celles associées aux multiples occurrences du lemme est conservée. La seconde, *-A-*, consiste à garder la moyenne des valeurs associées à toutes les occurrences d'un même lemme.

Les scores associés à chacun des lemmes des transcriptions automatiques peuvent également être calculés en combinant la pondération *tf-idf* et les scores acoustiques précédemment décrits. Cette combinaison est effectuée de la façon suivante :

$$S(w) = \frac{\theta_{ir}\ S_{ir}(w) + \theta_{ac}\ S_{ac}(w)}{\theta_{ir} + \theta_{ac}}\ , \tag{6.5}$$

avec $S_{ac}(w)$ le score acoustique obtenu grâce à la méthode décrite en section 3.2. Les deux facteurs θ_{ir} et θ_{ac} sont utilisés pour donner plus ou moins de poids aux différentes sources d'information.

Le tableau 6.1 présente des exemples de vecteurs caractéristiques pondérés par le critère *tf-idf*, un score acoustique (la valeur de l'énergie plus précisément) ou une combinaison des

TAB. 6.1 – Extraits des vecteurs caractéristiques pondérés par un score *tf-idf*, des informations prosodiques (*ac*) ou une combinaison des deux types d'information

tf-idf	*ac*	*tf-idf* + *ac*
degré 1	**chose 0.99**	degré 0.93
température 0.99	dérèglement 0.90	température 0.89
climatique 0.81	degré 0.87	climatique 0.80
océan 0.49	augmentation 0.85	océan 0.65
planète 0.39	océan 0.82	dérèglement 0.62
dérèglement 0.34	température 0.80	planète 0.52
augmentation 0.16	climatique 0.79	**chose 0.55**
chose 0.10	planète 0.65	augmentation 0.50

deux. Ces vecteurs ont été calculés à partir d'un segment extrait du journal télévisé du 2 février 2007 traitant du réchauffement climatique. Dans la colonne *ac*, nous constatons que le mot *augmentation*, souligné, est proéminent dans le discours, son score acoustique étant relativement élevé. Or ce mot, caractéristique du contenu du segment thématique, n'est pas associé à une pondération *tf-idf* importante, sa fréquence d'apparition étant probablement conséquente dans la collection de documents utilisée pour calculer les valeurs *idf*. Dans ce cas, la combinaison des deux indices, acoustique et textuel, permet au mot *augmentation* d'être considéré comme plus caractéristique du contenu du segment, reflétant ainsi l'intention du locuteur. Cependant, ce tableau montre également que la prosodie reflète divers phénomènes acoustiques qu'il est parfois difficile d'interpréter. Par exemple, le mot *chose*, en gras, est accentué par le locuteur mais ne correspond pas à un mot caractéristique du thème du réchauffement climatique.

Lorsque les mots clés caractéristiques utilisés pour la mise en relation des segments sont pondérés par les seuls scores acoustiques, nous avons pu constater que les différentes caractéristiques du signal employées avaient une influence variable sur les résultats obtenus. En effet, nous avons vu dans la section 3.2 que les informations acoustiques extraites du signal correspondent à l'*intensité*, représentant le niveau sonore perçu, au *pitch*, correspondant à la proéminence d'un mot dans le discours, ou à une combinaison des deux. Or, l'utilisation de ces trois informations acoustiques pour la mise en relation de segments thématiques a montré que la combinaison des deux indices, *intensité* et *pitch*, fournissait les meilleurs résultats, que ce soit sur le corpus de journaux télévisés ou sur le corpus composés des émissions *Sept à Huit*. De plus, nous avons remarqué que les quatre stratégies utilisées pour associer un score à chacun des lemmes de nos transcriptions[3] avaient un impact non négligeable sur la qualité des résultats. La stratégie MAX fournit les meilleurs résultats pour les deux corpus, suivie par la méthode ET, tandis que les performances obtenues grâce aux techniques MIN et MOYENNE sont les moins satisfaisantes, ce qui s'explique aisément par le fait que ces deux stratégies favorisent les mots associés aux valeurs de *pitch* ou d'*intensité* les plus faibles. Finalement, la gestion des répétitions de vocabulaire au sein d'une transcription -*M*- conduit à la meilleure mise en relation des segments, que ce soit pour les journaux télévisés ou les émissions de reportages.

[3]Une valeur de pitch ou d'intensité ayant été extraite pour chaque 0,01 seconde du signal, et la durée de prononciation d'un mot étant généralement supérieure à cette valeur, les quatre stratégies MAX, MOYENNE, MIN et ET consistent à calculer la valeur du score associé au lemme dans la transcription, *cf. 3.2.*

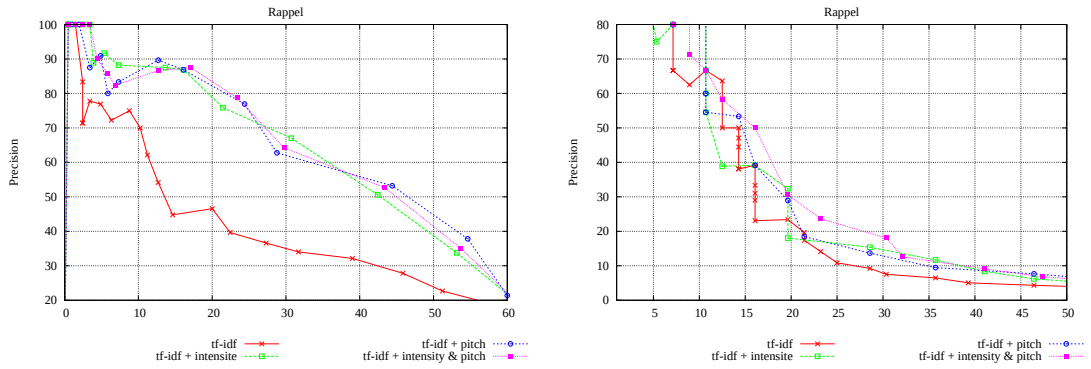
TAB. 6.2 – Premiers mots clés obtenus pour un reportage sur la disparition du petit Antoine

tf-idf	antoine alexandrie stéphane gendarme inspecter disparition restaurant maman enfant disparaître
informations acoustiques	rien couple stéphane soumettre moindre cas mot compagnon mystérieux gendarme

Si les meilleurs résultats sont obtenus pour la même combinaison de paramètres pour les deux corpora, l'utilisation d'informations acoustiques n'atteint pas les mêmes performances sur les émissions *Sept à Huit* que sur les *JT*. En effet, si la mise en relation des segments obtenue par le biais d'informations acoustiques est tout à fait comparable à celle obtenue grâce au critère *tf-idf* dans le cas des journaux télévisés, les performances observées pour les émissions de reportages sont largement inférieures. Cette différence s'explique selon nous principalement par le fait que les journalistes commentant des émissions de reportages adoptent, comme nous l'avons signalé précédemment, un ton très différent de celui employé dans les journaux télévisés. Or, dans ce cas, le fait qu'un mot est proéminent n'est pas toujours associé à la valeur informative de ce mot. Le tableau 6.2 présente les premiers mots clés correspondant à un reportage abordant la disparition d'un petit garçon, Antoine. Si les premiers mots fournis par le critère *tf-idf* reflètent bien le sujet abordé dans le segment, les termes obtenus grâce aux informations acoustiques, qui sont tous correctement reconnus par le système de transcription, ne traduisent pas le thème du reportage.

Les vecteurs caractéristiques des segments thématiquement cohérents peuvent également être composés de mots pondérés par une combinaison du critère *tf-idf* et des informations acoustiques. Cette combinaison des deux indices permet d'améliorer les performances du système de mise en relation des segments. En ce qui concerne les journaux télévisés, le gain le plus important est obtenu lorsque les paramètres θ_{ir} et θ_{ac} sont égaux à 1[4], c'est-à-dire lorsque les informations acoustiques et le critère *tf-idf* ont une influence comparable lors du calcul des vecteurs caractéristiques. L'amélioration observée, présentée dans la figure 6.2, montre ainsi que les informations prosodiques extraites constituent un indice important pour la mise en relation de segments thématiquement homogènes. Cette tendance a, par ailleurs, été confirmée lors de l'intégration de scores acoustiques calculés grâce à une méthode différente, nécessitant une phase d'apprentissage. Le calcul de ces scores ainsi que les résultats obtenus sont décrits dans (Guinaudeau and Hirschberg, 2011). Concernant les émissions de reportages *Sept à Huit*, nous pouvons constater sur la figure 6.2 que l'utilisation conjointe du critère *tf-idf* et des informations acoustiques ne permet pas d'améliorer les performances de façon aussi importante que pour les journaux télévisés. Cette combinaison est obtenue pour des valeurs de paramètres θ_{ir} et θ_{ac} égaux à 3 et 1 respectivement, ce qui traduit le fait que les scores acoustiques doivent avoir moins d'impact lors du calcul des vecteurs caractéristiques (une valeur de θ_{ac} plus importante entraîne une dégradation des performances, *cf.* annexe C). Ces différences entre les deux corpora, tant du point de vue des résultats que des valeurs de paramètres, ne sont cependant pas surprenantes au regard des résultats obtenus pour les scores acoustiques seuls.

[4]Voir l'annexe C pour plus de détail sur les résultats obtenus avec des valeurs de θ_{ir} et θ_{ac} différentes.



corpus *JT* corpus *Sept à Huit*

FIG. 6.2 – Combinaison du critère *tf-idf* et d'informations acoustiques pour la mise en relation de segments thématiquement homogènes extraits des corpora de journaux télévisés ou d'émissions de reportages *Sept à Huit*

6.2.2 Modification du calcul de la similarité entre vecteurs

Afin d'adapter des techniques développées pour du texte écrit aux particularités de nos données transcrites, nous avons modifié le calcul de la similarité entre vecteurs afin d'y introduire la prise en compte de relations sémantiques. Comme pour la segmentation thématique, l'utilisation de ces relations sert deux objectifs. Premièrement, elles doivent gérer le problème des synonymes en autorisant la mise en relation de deux segments thématiquement liés mais dont les vecteurs caractéristiques diffèrent de façon importante. Deuxièmement, l'intégration de relations sémantiques a pour but de pallier les erreurs de transcription en pénalisant les mots qui ne sont pas sémantiquement liés au reste du vocabulaire.

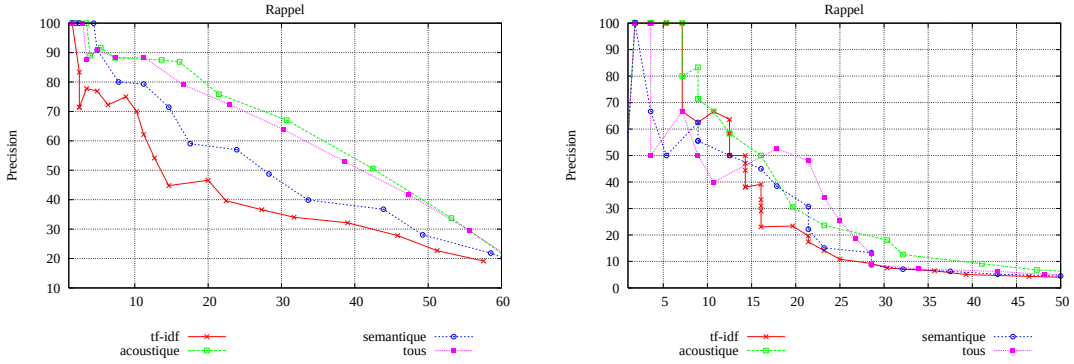
Les relations sémantiques utilisées dans ce travail sont des relations syntagmatiques et paradigmatiques extraites automatiquement d'un corpus composé d'article *du Monde* et de *l'Humanité* ainsi que des transcriptions manuelles des campagne ESTER 1 et ESTER 2. Nous avons choisi pour ces tâches de faire varier le nombre de relations introduites et d'employer les différentes techniques de sélection décrites dans la section 3.1.2.5.

Pour prendre en compte ces relations sémantiques, la mesure de similarité cosinus (6.4) a été modifiée de la façon suivante :

$$cos(A,B) = \frac{\sum_{j=1}^{L}(a_j + \sum_{k=1,k\neq j}^{L} r(a_j,b_k)) * (b_j + \sum_{k=1,k\neq j}^{L} r(b_j,a_k))}{\sqrt{\sum_{j=1}^{L}(a_j + \sum_{k=1,k\neq j}^{L} r(a_j,b_k))^2}\sqrt{\sum_{j=1}^{L}(b_j + \sum_{k=1,k\neq j}^{L} r(b_j,a_k))^2}} \quad , \qquad (6.6)$$

avec $r(a_j, b_k)$ la valeur de la proximité sémantique des mots a_j et b_k. Intuitivement, cette modification va permettre de prendre en compte les liens existants entre le mot a_j et les mots apparaissant dans B qui lui sont sémantiquement reliés.

La prise en compte de relations sémantiques lors du calcul de la similarité entre des segments thématiquement homogènes permet d'améliorer les performances du système lorsque le nombre de relations introduites n'est pas trop élevé. En effet, des expériences menées avec un nombre variable de relations introduites montrent que, au-delà d'un certain seuil, plus le nombre de relations augmente plus les résultats se dégradent (*cf.* annexe C pour plus de détail sur ces expériences). Nous avons également pu remarquer que les méthodes de sélection *ParMot* et *Total* influençaient toutes deux la qualité des mises en relation obtenues, la sélection



corpus *JT* corpus *Sept à Huit*

FIG. 6.3 – Combinaison du critère *tf-idf*, des informations acoustiques et des relations sémantiques pour la mise en relation de segments thématiquement homogènes extraits des corpora de journaux télévisés ou d'émissions de reportages *Sept à Huit*

ParMot offrant les meilleures performances. Finalement, nous avons constaté que les relations paradigmatiques étaient plus adaptées à notre tâche de mise en relation de segments que les relations syntagmatiques. Les relations paradigmatiques correspondant à des relations de synonymie ou d'hyperonymie, il n'est pas surprenant qu'elles permettent d'améliorer de façon plus importante les performances de notre système, comparées aux relations syntagmatiques qui regroupent les mots apparaissant couramment l'un au voisinage de l'autre. En comparant les résultats obtenus sur nos deux corpora de test (*cf.* figure 6.3) nous remarquons que l'intégration de relations sémantiques a plus d'influence sur le corpus de journaux télévisés que sur celui d'émissions de reportages. Cette différence entre les deux corpora est très certainement liée au fait que le nombre de mots différents est quatre fois plus important dans les reportages que dans les segments de JT. Ainsi, si la mise en relation de deux segments abordant des thématiques proches peut être problématique dans les journaux télévisés lorsque le vocabulaire est différent, ce cas de figure arrive plus rarement dans les émissions de reportages dans lesquelles le vocabulaire utilisé est plus varié. L'utilisation des relations sémantiques ayant également pour objectif de pallier les erreurs de transcription, comme suggéré dans (Johnson et al., 2000), nous avons comparé les résultats obtenus sur les transcriptions automatiques des émissions *Sept à Huit* avec ceux acquis à partir de transcriptions manuelles. L'intégration de relations paradigmatiques sélectionnées par la technique *ParMot*, qui fournit les meilleurs résultats pour les transcriptions automatiques, ne permet pas d'améliorer les performances de la mise en relation des segments lorsque celle-ci est effectuée à partir des transcriptions manuelles. Nous pouvons donc conclure de cette comparaison que les relations sémantiques introduites dans le corpus d'émissions de reportages *Sept à Huit* servent bien à pallier les erreurs de transcription.

L'utilisation des relations sémantiques peut également être combinée avec la prise en compte d'informations prosodiques pour tenter d'améliorer les performances du système de mise en relation de segments thématiques. Cependant, cette combinaison n'est pas réellement pertinente, que ce soit pour les journaux télévisés – pour lesquels les informations prosodiques associées au critère *tf-idf* fournissent les meilleurs résultats – ou pour les émissions de reportages (*cf.* figure 6.3).

6.3 Applications

Les étapes de segmentation thématique – présentée dans les chapitres 4 et 5 – et de mise en relation des segments permettent de développer des applications utiles pour la structuration automatique de flux télévisés. Nous avons pu, au cours de cette thèse, démontrer le bon comportement de nos méthodes à travers la réalisation de deux applications. La première, développée pour l'Institut National de l'Audiovisuel, est décrite en section 6.3.1. Elle consiste à associer aux notices documentaires décrivant le contenu de reportages de journaux télévisés, les frontières temporelles de ces reportages dans la vidéo. La seconde, présentée en section 6.3.2, propose une délinéarisation de flux télévisuels autorisant des utilisateurs à naviguer au sein d'une collection de documents.

6.3.1 Association de notices documentaires et de reportages télévisés

Dans le cadre du dépôt légal de la radio-télévision, l'Institut National de l'Audiovisuel (INA) archive près de 900 000 heures de télévision chaque année. Une partie de ces émissions est indexée et analysée par les documentalistes qui produisent un résumé ainsi qu'une indexation thématique de l'émission. Parmi ces émissions, les journaux télévisés sont segmentés en reportages, chaque reportage étant associé à une liste de mots clés, composée de noms communs et d'entités nommées, ainsi qu'au nom des journalistes auteurs du reportage, *etc.* Si ces notices permettent de décrire précisément le contenu de l'émission, elles sont cependant inadaptées à la recherche automatique d'un reportage particulier dans une collection de journaux télévisés. En effet, les informations concernant les temps de début et de fin des reportages considérés n'ayant pas été stockées dans les notices, il n'est pas possible, dans l'état actuel des choses, de faire le lien entre une notice et le segment correspondant dans le journal télévisé.

Afin d'associer des informations temporelles à chacune des notices documentaires produites pour dix journaux télévisés diffusés en février 1989, nous avons proposé de combiner notre méthode de segmentation thématique et notre technique de mise en relation de segments thématiquement homogènes. Chaque émission a ainsi été transcrite puis segmentée automatiquement afin d'extraire les différents reportages qui la compose. Ces reportages, ainsi que les différentes notices documentaires associées à l'émission, ont ensuite été représentés grâce à des vecteurs de mots clés pondérés par un critère *tf-idf*, comme décrit dans la section 6.1.2. Cependant, contrairement à ce qui été présenté en 6.1.2, où un même segment peut être associé à plusieurs autres, nous avons cherché, dans cette application, à mettre en relation chaque notice avec un seul reportage. Les éléments considérés comme reliés sémantiquement sont donc sélectionnés de façon à ce que, pour chaque notice, seul le couple notice-reportage associé à la valeur de similarité la plus élevée soit conservé. Nous avons également fait en sorte que deux notices ne puisse être finalement associées à un même reportage.

La figure 6.4 présente les résultats obtenus pour le journal télévisé diffusé le 1er février 1989 sur la chaîne de télévision France 2. Dans cette figure, la ligne du haut correspond à un alignement entre les notices et la vidéo obtenu manuellement et celle du bas représente l'alignement résultant de la combinaison des méthodes de segmentation et de mise en relation des segments thématiques. Chaque reportage est symbolisé par un rectangle à la couleur et au motif variable. Nous pouvons constater, sur cette figure, que bien que les frontières temporelles ne soient pas parfaites, chaque notice est associée au reportage qu'elle décrit. La qualité des résultats obtenus dépend de deux éléments. Tout d'abord, les erreurs de segmentation

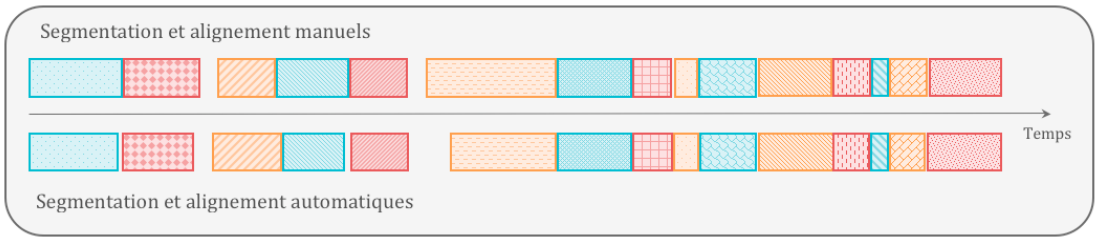


FIG. 6.4 – Association automatique des notices documentaires aux reportages d'un journal télévisé

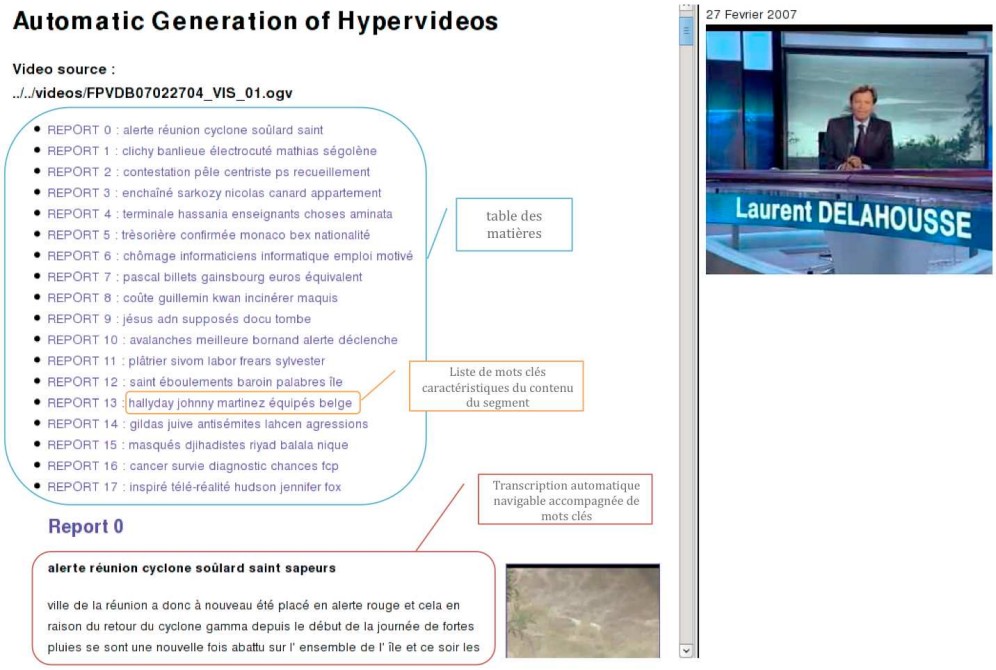
thématique peuvent avoir deux types de conséquences : si un segment contient plusieurs reportages – erreur de sous-segmentation – alors une notice peut ne pas être associée à son reportage, au contraire, si un reportage est découpé en plusieurs segments – erreur de sur-segmentation – la notice correspondante est associée à une seule partie du reportage et ses frontières de début et de fin ne correspondent pas tout à fait à la durée du reportage. De plus, la nature des notices produites par les documentalistes de l'INA a également un impact sur la qualité de l'alignement. En effet, lorsque les documentalistes de l'INA utilisent les termes prononcés dans le reportage pour décrire son contenu, la mise en relation des notices et des reportages se trouve facilitée. Cependant, dès que la qualité de ces notices diminue, les performances de l'application proposée se dégradent.

6.3.2 Délinéarisation de flux télévisuels

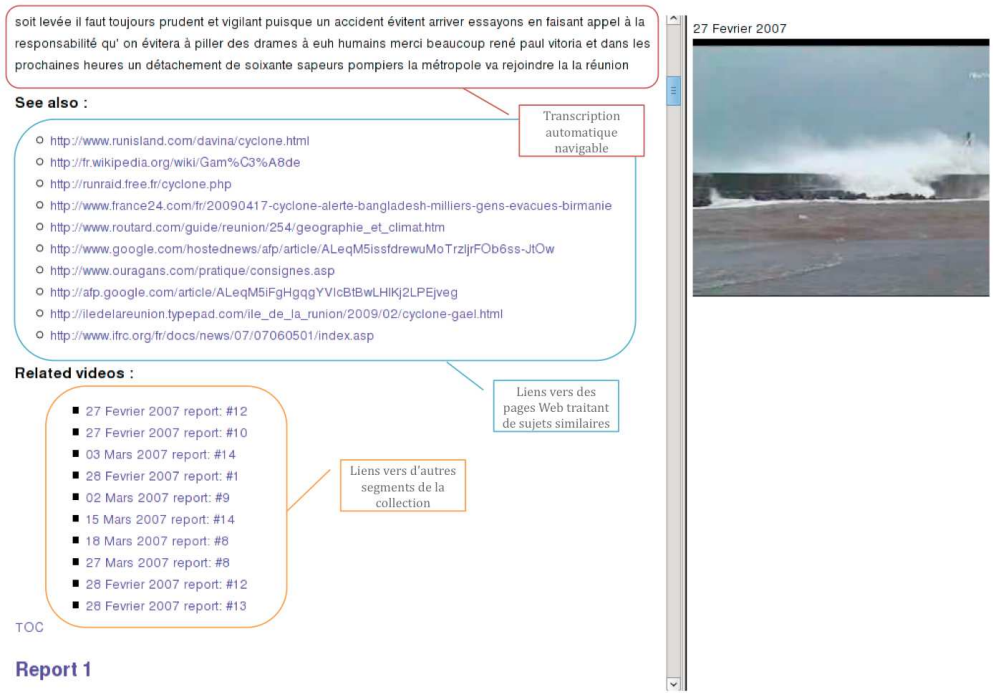
La combinaison des méthodes développées dans cette thèse pour la structuration thématique linéaire ont également permis la mise en place d'une application de délinéarisation de flux télévisuels, présentée lors du NEM Summit 2009[5] (*cf.* (Gravier et al., 2011) pour plus de détail sur les techniques mises en œuvre). Dans ce cadre, les journaux télévisés de notre corpus ont été segmentés grâce à la méthode présentée dans les chapitres 4 et 5. La similarité entre les vecteurs caractéristiques de ces segments a ensuite été calculée afin de mettre en relation les segments abordant des faits d'actualité similaires.

La figure 6.5(a) présente la table des matières du journal télévisé du 27 février 2007, obtenue automatiquement grâce à la segmentation thématique et à l'extraction des mots clés caractéristiques de chaque segment. Chaque reportage extrait du JT est associé à sa transcription automatique (encadré rouge) ainsi qu'à la vidéo correspondante. La figure 6.5(b) montre les liens créés entre un reportage traitant de l'arrivée du cyclone Gamède sur l'île de la Réunion et les différents reportages de la collection (encadré orange) qui abordent des sujets similaires. Ces reportages peuvent appartenir à la même émission, comme pour les deux premiers éléments – qui traitent du même sujet ou d'autres catastrophes naturelles – ou apparaître dans d'autres journaux télévisés. Dans ce cas, les reportages concernés abordent les conséquences et dégâts liés au cyclone, les risques d'épidémies, *etc.* La figure 6.5(b) présente également (encadré bleu) une liste de liens pointant vers des pages Web abordant des sujets similaires. Ces liens sont obtenus automatiquement, par le biais des mots clés caractéristiques du reportage, grâce à une technique développée par Lecorvé (Lecorvé et al., 2008).

[5]Nous tenons à remercier Mathieu Ben, Sébastien Campion et Gwénolé Lecorvé pour leurs contributions au sein de ce travail collaboratif.



(a) Table des matières du journal télévisé diffusé le 27 février 2007 sur France 2. Chaque segment est caractérisé par une liste de mot clés



(b) Présentation des liens pointant vers d'autres reportages de la même collection de documents et vers des pages Web abordant des sujets similaires

FIG. 6.5 – Capture d'écran de la démonstration de délinéarisation de flux télévisuels présentée au Nem Summit 2009

6.4 Bilan du chapitre

Au sein de la problématique de la structuration thématique linéaire, ce chapitre s'est intéressé à la mise en relation de segments thématiquement homogènes obtenus lors d'une phase préalable de segmentation thématique. Cette mise en relation a été adaptée à nos données télévisuelles grâce à l'utilisation de deux indices.

Premièrement, nous avons pris en compte le fait que les documents sur lesquels nous travaillons ne se résument pas aux transcriptions automatiques fournies par le système de reconnaissance automatique de la parole. La dimension orale de nos données constitue, en effet, un indice important dans la mise en relation de segments thématiques. Nous avons pu montrer, par exemple, que la combinaison d'un indice textuel, le critère *tf-idf*, et d'informations acoustiques permet d'améliorer les performances du système, que ce soit pour le corpus de journaux télévisés ou, dans une moindre mesure, pour le corpus d'émissions de reportages. Cette constatation est intéressante car, si l'apport de la prosodie a été démontré dans les systèmes de recherche dans les documents oraux pour des langues très accentuées comme le chinois et l'anglais, aucune étude n'avait permis de mettre en avant, à notre connaissance, son intérêt pour la langue française.

Deuxièmement, nous avons intégré des relations sémantiques lors du calcul de la similarité entre les segments thématiques. Si l'utilisation d'un tel indice n'est pas nouveau en soi, la comparaison des résultats obtenus sur des transcriptions automatiques et manuelles nous a permis de montrer que la conclusion proposée par (Johnson et al., 2000) pouvait s'étendre à des relations obtenues automatiquement.

Au cours de ce chapitre, nous avons également pu démontrer le bon fonctionnement des méthodes développées, que ce soit pour la tâche de segmentation thématique ou la tâche de mise en relation des segments, grâce à la mise en place de deux applications de structuration thématique linéaire.

Afin d'améliorer les performances du système de mise en relation de segments audiovisuels thématiquement homogènes, deux pistes principales peuvent être étudiées. Premièrement, l'utilisation de la prosodie fournissant un gain important dans les résultats obtenus, il nous semble intéressant de travailler sur des méthodes d'extraction plus sophistiquées. En effet, l'obtention des informations prosodiques utilisées dans ces travaux repose sur la simple extraction de deux caractéristiques audios du signal, l'*intensité* et le *pitch*. Or il existe d'autres indices permettant de traduire les phénomènes prosodiques avec plus de précision. Par exemple, le *pitch*, qui correspond à la proéminence d'un mot dans le discours, peut être associé à cinq types d'accents dans la langue anglaise, deux simples et trois complexes. Si l'anglais est une langue plus accentuée que la langue française, il est cependant probable qu'il existe des subtilités dans l'accentuation du français qui pourraient être prise en compte grâce à ces différents accents. De plus, la normalisation des informations acoustiques extraites pourrait être améliorée. En effet, comme nous l'avons précisé dans le chapitre 3, la normalisation par locuteur a été effectuée à partir d'une segmentation en locuteurs fournie par le système de transcription. Ce découpage n'étant pas optimal, la normalisation obtenue reste approximative. L'amélioration des performances du système nécessite également de fournir un effort concernant la gestion des erreurs de transcription, ces erreurs ayant un impact sur la qualité de la mise en relation. Pour ce faire, il pourrait être intéressant d'utiliser les différentes sorties fournies par le système de reconnaissance automatique de la parole lors de la représentation des segments. L'utilisation des mesures de confiance pour modifier la pondération *tf-idf* est la piste qui nous semble la plus évidente à mettre en œuvre pour diminuer

le poids des mots mal transcrits. Cependant, l'utilisation des mesures de confiance n'est pas immédiate, de premières expériences consistant à modifier le poids *tf-idf* des mots en fonction de leur mesure de confiance n'ayant pas permis d'améliorer les performances du système. En effet, les mesures de confiance n'étant pas toujours fiables, certains mots mal transcrits ont vu leur pondération *tf-idf* augmentée, bruitant ainsi la mise en relation des segments. La prise en compte des différentes hypothèses de transcription considérées par le système peut également être envisagée pour améliorer le calcul de la similarité des segments. En effet, un segment contenant un mot w mal transcrit dans la transcription finale peut cependant être associé à un graphe de mots contenant la transcription correcte de w qui n'a pas été retenue pour apparaître dans la sortie finale. Les relations sémantiques employées ayant montré leur utilité pour gérer le faible taux de répétition dans le vocabulaire de nos données et pour pallier les erreurs de transcriptions, il nous semble également intéressant de chercher à obtenir des relations plus adaptée à nos données. Dans (Lecorvé et al., 2008), Lecorvé et al. ont proposé un système permettant d'extraire automatiquement d'internet un corpus thématique à partir de segments thématiquement homogène. Ce corpus peut être, selon nous, utilisé pour calculer des relations sémantiques liées au thème des segments à mettre en relation, ce qui devrait augmenter l'influence des relations sémantiques lors de la mise en relation. Finalement, un dernier élément que nous n'avons pas pris en compte dans ce travail et qui pénalise, selon nous, la mise en relation des segments thématiquement homogènes est la présence, dans la parole contenue dans les segments, de mots qui n'apparaissent pas dans le vocabulaire du système de transcription. Or ces mots, souvent des entités nommées, sont généralement très caractéristiques de l'événement décrit dans le reportage. Pour prendre en compte ces mots hors vocabulaire, certains travaux de *Spoken Document Retrieval* proposent de passer par une représentation phonétique de la parole prononcée et d'opérer une comparaison entre ces représentations (Witbrock and Hauptmann, 1997; Logan et al., 2002). Des expériences menées sur la mise en relations de programmes télévisuels et de leur description textuelle extraite d'un guide de programmes ayant montré que l'utilisation de cette représentation phonétique entraînait beaucoup de bruit (*cf.* (Guinaudeau et al., 2009)), cette dernière piste doit toutefois être, selon nous, étudiée avec prudence.

Chapitre 7

Structuration thématique hiérarchique

La structuration proposée dans le chapitre précédent permet d'organiser des (parties) de documents thématiquement homogènes les uns par rapport aux autres et de produire, ainsi, la structuration d'une collection de documents. La structuration peut également être réalisée à l'intérieur même du document, par le biais d'une structuration hiérarchique, pour mettre en avant les différents niveaux de granularité qui structurent l'information présente dans le document. Cette structuration hiérarchique peu également être utile en recherche d'information (Hearst and Plaunt, 1993), où elle peut permettre de juger de la pertinence d'un document à plusieurs niveaux de détail, et en résumé automatique (Angheluta et al., 2002; Li et al., 2001).

Dans le cadre de la structuration automatique de documents télévisuels, de nombreux travaux ont proposé des techniques permettant d'extraire l'organisation interne d'une émission. C'est le cas, par exemple, de (Kijak, 2003) et (Delakis, 2006) qui développent des méthodes d'identification des phases de jeux dans des matches de tennis, fournissant ainsi aux utilisateurs une table des matières de ces programmes. De façon analogue, les travaux de segmentation thématique de documents textuels ont mis en évidence la nécessité de travailler sur une structuration hiérarchique des documents. En effet, nombre de travaux s'accordent à dire que la structure d'un texte est plus complexe qu'une simple organisation linéaire. Mann et Thompson (Mann and Thompson, 1988) proposent, par exemple, une théorie sur les aspects majeurs de la structure rhétorique de documents textuels. De nombreuses théories du discours, sur le modèle de (Grosz and Sidner, 1986), considèrent également que les documents textuels possèdent une structure dans laquelle les segments thématiques ne s'organisent plus de façon consécutive mais de façon hiérarchique ; un sujet abordé dans un segment peut être découpé en plusieurs sous-sujets et éventuellement englobé dans un cadre de discours plus général. Cependant, ces différents niveaux de granularité sont souvent ignorés lors de la structuration de documents qui se limite généralement à une segmentation linéaire. Or, s'ils sont à l'origine des difficultés de mise en place de segmentation thématique linéaire de référence – problème d'accord inter-annotateurs sur la granularité de la segmentation à proposer –, ils permettent également de traduire de façon plus riche l'organisation interne du document.

Comme pour les techniques de structuration présentées plus tôt dans ce manuscrit, la méthode de *structuration thématique hiérarchique* décrite dans ce chapitre se base sur les transcriptions automatiques de la parole prononcée durant les émissions. Afin d'extraire l'organisation hiérarchique d'un document télévisuel, nous souhaitons employer sur nos données transcrites une technique de segmentation thématique hiérarchique. Le nombre de travaux abordant le problème de la segmentation thématique hiérarchique étant relativement faible,

nous proposons dans ce chapitre un travail exploratoire visant à examiner différentes pistes permettant de passer d'une segmentation linéaire à une structuration thématique hiérarchique. Pour cela, nous présentons tout d'abord les différentes méthodes de segmentation hiérarchique existantes, les méthodes d'évaluation utilisées ainsi que les quelques définitions de la granularité au sein de la notion de thème. Nous décrivons ensuite les pistes que nous avons explorées pour adapter l'algorithme de segmentation thématique linéaire employé dans les chapitres précédents à la tâche de segmentation hiérarchique et les résultats obtenus. Finalement, nous donnons différentes voies d'amélioration des techniques proposées, ainsi que des perspectives à plus long terme.

7.1 Segmentation thématique hiérarchique : principe et état de l'art

Comparativement à la segmentation thématique linéaire, peu de travaux de traitement automatique des langues se sont penchés sur la problématique de l'organisation thématique hiérarchique de documents textuels. Dans cette partie, nous exposons, tout d'abord, les notions de thème et de sous-thèmes telles qu'elles sont abordées dans la littérature et dans le cadre plus précis de données télévisuelles. Nous présentons, également, une vue d'ensemble des différentes techniques existantes et positionnons les pistes développées dans cette thèse par rapport à ces travaux. Finalement, nous abordons, dans la section 7.1.3, la problématique de l'évaluation des algorithmes de segmentation thématique hiérarchique.

7.1.1 Hiérarchie au sein des thèmes

Si le concept de thème a fait l'objet de nombreuses définitions (*cf.* section 4.1), la notion de granularité ou de hiérarchie au sein des thèmes a été comparativement très peu abordée dans la littérature linguistique. Dans (Rastier, 1987), Rastier propose une hiérarchisation domaine/taxème/sémème pouvant s'apparenter à cette notion de granularité. Pour Rastier, les domaines sont en effet des classes, comprenant des sémèmes ayant un trait générique, qui correspondent à une pratique sociale alors que les taxèmes sont le reflet de situations de choix dans des pratiques concrètes ou théoriques. Par exemple, les sémèmes {/métro/, /train/, /autobus/, /autocar/} relèvent du domaine //transports// articulé en deux taxèmes, le taxème //ferré//, comprenant les sémèmes /métro/ et /train/, et le taxème //routier//, contenant les sémèmes /autobus/ et /autocar/.

Nous pouvons, également, trouver des formalisations de la notion de hiérarchique au sein des thèmes dans des travaux de traitement automatique des langues ou dans les tâches de *Topic Detection and Tracking*. Dans le cadre d'un travail de structuration de discours (Choi, 2000b), Choi différencie la notion de thème de celle de sous-thème de la façon suivante : selon lui, un segment thématiquement cohérent, considéré comme une séquence de phrases interdépendantes, caractérise un sous-thème si son interprétation est dépendante d'un autre segment thématique. Les différents niveaux de granularité peuvent aussi être vus comme les parties structurant un document textuel. Dans (Slaney and Ponceleon, 2001), les auteurs proposent ainsi de développer une technique de segmentation thématique et testent sa capacité à retrouver l'organisation en sections et sous-sections d'un chapitre de livre. De la même manière, les exemples de vérité terrain présentés dans (Hsueh et al., 2006), développant une méthode de segmentation hiérarchique de transcriptions automatiques de réunions, s'apparentent aux

différents éléments composant un article scientifique. Les organisateurs du projet *Topic Detection and Tracking* (TDT) ont également proposé des définitions des notions de *topic* et d'*event* qui peuvent être considérées comme traduisant des niveaux de granularité différents au sein d'un fait d'actualité. Pour (Fukumoto and Suzuki, 2000), la différence entre un *topic* et un *event* est cependant difficile à faire ce qui les conduit à proposer la différenciation suivante. Un *event* fait référence au sujet du fait d'actualité, c'est-à-dire ce que l'auteur veut exprimer à propos de ce fait, en d'autres termes les notions de qui, quoi, où, quand, pourquoi et comment. Le *topic* représente, quant à lui, un élement unique, ayant lieu à une date et une place précises, ainsi que ces conséquences. Le *topic* est donc considéré comme l'origine, l'arrière-plan de l'événement. De ce fait, il existe dans la définition du concept de l'*event* une notion d'évolution qui n'est pas présente dans la définition du *topic*.

La notion de granularité au sein de journaux télévisés ne peut cependant pas se limiter à une notion d'évolution ou à une dimension chronologique. Dans le cadre de notre travail de structuration thématique hiérarchique de documents télévisuels, nous proposons une granularité composée de deux niveaux de hiérarchie dans laquelle les thèmes et sous-thèmes sont considérés de la façon suivante. Le thème correspond à un reportage associé à ses plateaux de lancement et de fin, les sous-thèmes représentant, quant à eux, les différents points de vue ou aspects de la problématique abordée dans le reportage. Les émissions de reportages *Envoyé Spécial* utilisées comme corpus de test possédant une structure hiérarchique assez claire, la différenciation entre thème et sous-thème est assez facilement identifiable[1]. Par exemple, le premier reportage de l'émission *Envoyé Spécial* diffusé le 15 janvier 2009, intitulé « La France de la débrouille », aborde le sujet de la crise financière en France et ses conséquences sur le budget des français. Au cours de ce reportage, les journalistes suivent le quotidien de six individus qui mettent en place différentes techniques permettant de consommer à bas coût. Le thème « La France de la débrouille » est donc divisé en neuf sous-thèmes : les six sous-thèmes correspondant aux témoignages de Marine, Toufik, Sabrina, Thibault, Mouna et Hassan, la conclusion ainsi que deux parties introductives – une générale et l'autre plus ciblée sur les six personnages suivis lors du reportage.

7.1.2 Segmentation thématique hiérarchique : état de l'art et positionnement

Si de nombreuses études s'accordent à dire que la structure d'un document n'est pas seulement linéaire, la segmentation thématique hiérarchique a été assez peu étudiée, comparativement à la segmentation thématique linéaire. Nous présentons dans cette section les quelques travaux dédiés au développement de méthodes de segmentation thématique hiérarchique. Nous décrivons également l'approche utilisée pour mener à bien notre objectif de *structuration thématique hiérarchique*.

7.1.2.1 État de l'art

Certaines techniques de segmentation thématique linéaire, (Utiyama and Isahara, 2001) par exemple, proposent d'extraire la structure hiérarchique des documents en appliquant de façon répétitive l'algorithme de segmentation sur les segments thématiquement homogènes obtenus lors d'une première itération. D'autres études, toutes très récentes, développent des

[1]La définition d'un troisième niveau de hiérarchie étant un peu plus problématique, nous avons choisi de limiter notre analyse à une structuration en deux niveaux.

techniques totalement consacrées à la segmentation thématique hiérarchique. Tous ces travaux utilisent des indices lexicaux et, plus précisément, le critère de cohésion lexicale. (Moens and Busser, 2001) et (Eisenstein, 2009) considèrent que la distribution des mots au sein du texte est un indice très important pour l'extraction d'une structure hiérarchique. Dans (Moens and Busser, 2001), cette distribution des mots est matérialisée par des chaînes lexicales qui relient les différentes occurrences des mots pleins du texte, les mots qui leur sont sémantiquement reliés (obtenus grâce à WordNet (Miller, 1995)), ainsi que leurs représentants référentiels dans le texte. Ces chaînes lexicales ne sont pas utilisées pour opérer la segmentation thématique en tant que telle[2] mais pour représenter les liens existants entre les différents segments : lien hiérarchique ou *retour sémantique*, c'est-à-dire la reprise d'un thème suspendu précédemment dans le texte. Contrairement à la méthode *bottom-up* mise en place dans (Moens and Busser, 2001) qui consiste à segmenter linéairement puis à mettre en relation les segments thématiquement homogènes présentant un lien hiérarchique, Eisenstein propose dans (Eisenstein, 2009) une méthode globale au cours de laquelle la segmentation hiérarchique est mise en place directement. Dans cet article, la segmentation hiérarchique est formalisée dans un cadre bayésien et repose sur l'hypothèse que chaque mot du texte w_t est représenté par un modèle de langue estimé sur une portion plus ou moins importante du texte. Ainsi un modèle de langue calculé sur le texte entier sera plus à même de représenter les mots qui apparaissent tout au long du texte alors qu'un modèle de langue calculé sur une section du texte plus restreinte sera plus pertinent pour expliquer des termes[3] qui apparaissent très localement dans le texte. Lors du calcul de la segmentation hiérarchique, l'algorithme cherche à maximiser la cohésion lexicale, grâce à un principe similaire à celui développé dans UI, des segments à chaque niveau de hiérarchique tout en obligeant les frontières de niveaux hiérarchique supérieurs à être alignées sur celles proposées lors des segmentations des niveaux inférieurs[4].

La segmentation thématique hiérarchique peut également être vue comme une tâche de *clustering*. En effet, comme le suggère Yaari (Yaari, 1997), une technique de *clustering* agglomérative hiérarchique utilisée pour la segmentation thématique linéaire semble exploitable pour extraire une segmentation thématique hiérarchique. Dans (Carroll, 2010), l'auteur adapte l'algorithme C99 de Choi (Choi, 2000a) afin d'extraire une organisation hiérarchique de la segmentation. Pour cela, il conserve une trace de l'ordre de séparation des segments lors de la phase de segmentation linéaire, pour extraire un ordonnancement de l'importance des frontières et ainsi différencier les frontières fortes définissant les thèmes des frontières plus faibles apparaissant entre les sous-thèmes.

Finalement, la dernière étude proposant une technique de segmentation thématique hiérarchique combine une méthode de traitement automatique des langues – l'indexation sémantique latente (LSI) – avec une technique de traitement du signal – la *scale-space segmentation* (Slaney and Ponceleon, 2001). Dans un premier temps, les documents sont représentés sous forme de matrices grâce à la LSI, chaque phrase S_i du texte étant représentée par un vecteur $\vec{H}(S_i)$ correspondant à la i^e colonne de la matrice. Une décomposition en valeurs singulières permet de réduire le nombre de dimensions à k (ici $k = 10$) en ne conservant que les k mots les plus importants. À partir de cette représentation, les auteurs appliquent ensuite la méthode de *scale-space segmentation* qui consiste à lisser chacune des dimensions des vecteurs de façon indépendante. Ce lissage est effectué grâce à un noyau gaussien associé à plusieurs valeurs

[2]Selon les auteurs, les chaînes lexicales ne représentent pas correctement les différentes thématiques apparaissant dans le document et elles peuvent se chevaucher au sein d'un même thème.

[3]*terme* est employé dans ce chapitre comme un synonyme de *mot* et non dans un cadre de terminologie.

[4]Le code source de cet algorithme est disponible à l'adresse http ://people.csail.mit.edu/jacobe/naacl09.html.

d'échelle σ. Pour chaque vecteur, le nombre de lissages existant entre la plus large échelle et l'échelle 0 traduit l'importance de la frontière thématique ainsi définie. C'est la différence d'importance entre les frontières qui va déterminer l'aspect hiérarchique de la segmentation.

7.1.2.2 Positionnement et approche retenue

Parmi les approches présentées dans la section précédente, la plus aboutie est celle proposée par Eisenstein qui permet d'obtenir une segmentation hiérarchique sans passer par une segmentation linéaire préalable. Cependant, cette méthode ne répond pas à notre souci de généricité ni à l'objectif de structuration sans supervision que nous nous somme fixé dans le cadre de cette thèse. L'algorithme proposé nécessite, en effet, un paramètre d'entrée spécifiant les durées attendues des segments et ce, à tous les niveaux de hiérarchie désirés. De plus, une analyse qualitative des résultats obtenus pour la segmentation en deux niveaux de hiérarchie d'une émission d'*Envoyé Spécial* a mis à jour l'incapacité de cet algorithme à traiter nos données. En effet, si les segments obtenus pour le niveau hiérarchique le plus élevé correspondent à la segmentation de référence, les sous-segments retournés sont extrêmement réguliers dans leur durée, ce qui ne correspond absolument pas à la réalité de notre corpus.

Si cette méthode ne semble pas applicable dans le cadre de notre travail, elle soulève cependant un point qui nous paraît très important, celui du positionnement du vocabulaire au sein du document à segmenter. Eisenstein remarque en effet que les mots caractéristiques d'un sous-segment apparaissent de façon très localisée dans le texte alors que ceux représentatifs des segments de hiérarchie supérieure ont une fenêtre d'apparition plus étendue. Les auteurs de (Moens and Busser, 2001) se basent également sur cette observation pour construire une segmentation hiérarchique à partir d'une segmentation linéaire par le biais de chaînes lexicales.

La problématique de la localisation du vocabulaire au sein du document à segmenter nous semble donc être un point crucial pour la segmentation thématique hiérarchique. Afin de mettre en place une segmentation thématique hiérarchique, nous avons choisi d'appliquer l'algorithme de Utiyama et Isahara, utilisé jusqu'ici dans le cadre de la segmentation thématique linéaire, de façon répétitive, c'est-à-dire en appliquant une nouvelle fois l'algorithme sur des segments thématiquement homogènes obtenus lors d'une première itération. Or, l'utilisation répétitive de l'algorithme n'est pas immédiate, les mots ayant participé à la cohésion lexicale lors de la segmentation des niveaux hiérarchiques supérieurs allant, de nouveau, jouer un rôle dans le calcul de la segmentation thématique des niveaux inférieurs. Afin de pénaliser les termes pris en compte lors du calcul de la cohésion lexicale lors des segmentations de niveau supérieur, nous proposons de modifier la méthode d'estimation de cette cohésion en nous basant sur la distribution du vocabulaire, comme le suggère (Eisenstein, 2009) et (Moens and Busser, 2001).

7.1.3 Évaluation de la segmentation thématique hiérarchique

L'évaluation de la segmentation thématique hiérarchique se heurte aux mêmes problèmes que celle de la segmentation thématique linéaire : difficultés à produire des segmentations de référence, problèmes de définitions de métriques, *etc.* De fait, la plupart des travaux présentés dans la section 7.1.2.1 ne mettent pas en place d'évaluation quantitative. Dans (Slaney and Ponceleon, 2001), les auteurs évaluent qualitativement leurs résultats en proposant deux exemples de sortie de leur algorithme, alors que la segmentation présentée dans (Moens and Busser, 2001) est utilisée dans un système de résumé automatique et est évaluée de façon indirecte.

Seul (Carroll, 2010) propose une méthode d'évaluation dédiée à la segmentation thématique hiérarchique à partir d'une extension de la mesure P_k. Cette mesure calcule, pour chaque niveau de hiérarchie i, la valeur de la mesure P_k entre les frontières de référence et les frontières hypothèses de niveau hiérarchique supérieur ou égal à i. Elle est définie de la façon suivante :

$$E_{P_k} = \frac{1}{|R|} \sum_i c_i P_k(R_i, H_i) \ , \tag{7.1}$$

avec R_i (resp. H_i) l'ensemble des frontières de référence (resp. hypothèses) de niveau supérieur ou égal à i. La mesure P_k est calculée pour chaque niveau de hiérarchie existant dans la segmentation de référence, c_i étant égal au nombre de frontières existant au niveau de hiérarchie i. De plus, lors du calcul de cette mesure, Carroll contraint $|R_i|$ a être égal à $|H_i|$ afin de s'adapter aux segmentations qui sous-estiment ou sur-estiment le nombre de segments. Si le nombre de frontières dans l'hypothèse est inférieur au nombre de frontières dans la référence, il propose de prendre en compte des frontières de rang inférieur (ou, pour l'évaluation des niveaux hiérarchiques très inférieurs, des frontières non marquées comme frontières qui peuvent être considérées comme des frontières de rang très inférieur). Cette contrainte posée sur l'égalité du nombre de frontières entre l'hypothèse et la référence nous semble dangereuse car elle ne permet pas, selon nous, de traduire le comportement réel d'un algorithme de segmentation, les comportements de sur- et de sous-segmentations étant fortement caractéristiques de la qualité de la segmentation produite.

Une autre possibilité existante pour évaluer les résultats d'une segmentation thématique hiérarchique consiste à utiliser la mesure proposée dans (Mohri et al., 2009). Cette mesure prend en effet en compte l'importance des frontières lors de l'évaluation. Ici l'importance d'une frontière, ajoutée ou supprimée à tort, est calculée à partir du contenu des segments considérés comme thématiquement cohérents. Par exemple, soit une segmentation de référence composée de trois segments $z_{1,r}$, $z_{2,r}$ et $z_{3,r}$, avec $z_{1,r}$ et $z_{2,r}$ ayant une distribution des mots assez similaires et $z_{2,r}$ et $z_{3,r}$ ayant une distribution des mots très différentes. La segmentation hypothèse H_1 manquant la frontière entre $z_{1,r}$ et $z_{2,r}$ doit être moins pénalisée que la segmentation hypothèse H_2 qui oublie une frontière entre $z_{2,r}$ et $z_{3,r}$. Si cette mesure n'a pas été définie spécifiquement pour la segmentation hiérarchique, elle peut tout à fait être employée dans ce cadre car elle permet de pénaliser de façon moins importante les frontières correspondant à une segmentation de granularité plus fine que les frontières correspondant à une segmentation de niveau hiérarchique supérieur. Cependant, cette technique d'évaluation repose sur la comparaison du vocabulaire existant entre les segments de la référence et ceux de la segmentation hypothèse, comparaison qui peut souffrir de la présence d'erreurs de transcription.

Finalement, la façon la plus élémentaire d'évaluer une segmentation hiérarchique, que nous avons retenue, est de produire une vérité terrain pour chaque niveau de hiérarchie et de comparer ensuite les segmentations hypothèses aux segmentations de référence pour chaque niveau de hiérarchie. Nous sommes consciente que cette méthode présente plusieurs inconvénients. D'une part, ce type d'évaluation *en cascade* ne va pas permettre de prendre en compte de façon globale une erreur qui se serait produite lors de la segmentation hiérarchique supérieure et qui va se répercuter sur les segmentations aux niveaux hiérarchiques inférieurs. D'autre part, la création de segmentations de référence à chaque niveau est une source de difficultés. En effet, si la création de référence pour une segmentation linéaire se heurte au problème de l'accord inter-annotateurs, le désaccord va être plus marqué encore lorsque les changements de thèmes sont plus subtils. Il conviendra donc, dans le cadre d'un travail plus poussé, d'étudier

TAB. 7.1 – Nombre d'occurrences des mots caractéristiques de trois sous-segments de « La France de la débrouille » dans les sous-segments considérés, le segment thématique de niveau hiérarchique supérieur et la totalité de l'émission *Envoyé Spécial*

Thème du sous-segment	mots apparaissant dans les sous-segments	nb occurrences dans le sous-segment	nb occurrences dans le segment	nb occurrences dans la l'émission complète
introduction	crise	3	6	9
	consommation	2	3	3
	reportage	2	2	6
Toufik	fruit	5	5	6
	légume	4	4	4
	rungis	2	2	3
Hassan	hassan	6	7	7
	soirée	5	5	9
	soir	4	5	16

cet aspect de la question en utilisant, par exemple, des comparaisons de documents structurés comme nous le proposons dans la section 7.3.

7.2 Segmentation hiérarchique de programmes TV

Comme nous l'avons spécifié dans la section 7.1.2.2, notre méthode de segmentation thématique hiérarchique consiste à utiliser l'algorithme de segmentation thématique linéaire, présenté dans le chapitre 4, de façon répétitive tout en prenant en compte la distribution des mots du vocabulaire au sein du document. Nous extrayons donc automatiquement des segments thématiquement homogènes – que nous appellerons segments de niveau hiérarchique supérieur – de nos émissions *Envoyé Spécial*, puis nous appliquons sur ces segments une version modifiée de l'algorithme de segmentation. Pour définir cette nouvelle version, nous examinons deux stratégies, décrites respectivement dans les sections 7.2.1 et 7.2.2 : une modification du calcul de la probabilité généralisée qui permet d'évaluer la cohésion lexicale d'un segment en pénalisant les mots ayant déjà fortement participé à l'estimation de la cohésion des segments de niveau hiérarchique supérieur d'une part, et l'utilisation de chaînes lexicales pour favoriser les termes apparaissant très localement dans les sous-segments thématiques d'autre part.

7.2.1 Modification de la probabilité généralisée

Les mots que nous souhaitons mettre en avant lors de la cohésion lexicale des sous-segments sont ceux qui apparaissent presque exclusivement au sein de chaque sous-segment. Le tableau 7.1 présente le nombre d'occurrences de termes représentatifs des thèmes abordés dans trois sous-segments de « La France de la débrouille ». Ce tableau nous montre que ces mots ont une fréquence d'apparition importante dans le sous-segment comparativement à leur nombre d'occurrences dans le segment thématique de niveau hiérarchique supérieur. Certains mots apparaissent, en effet, exclusivement dans le sous-segment qu'ils caractérisent. C'est le cas, par exemple, de *légume* dont toutes les occurrences se situent au sein du sous-segment présentant l'histoire de Toufik qui achète des fruits et légumes en gros à Rungis.

TAB. 7.2 – Valeurs des scores obtenus pour des mots apparaissant k fois dans un sous-segment et K fois dans un segment de niveau hiérarchique supérieur.

	k	K	probabilité généralisée baseline	probabilité généralisée normalisée
u	5	5	-5,47	0,71
v	5	7	-5,47	0,37

Dans le cadre de la segmentation thématique hiérarchique, la probabilité généralisée se calcule de la façon suivante :

$$\ln P[S_{ij};\Delta_{ij}] = \sum_{w \in S_i} n_{ij}(w) \ln P[w;\Delta_{ij}] \ , \tag{7.2}$$

avec S_{ij} le j^e sous-segment du segment S_i obtenu lors d'une première itération de l'algorithme. $n_{ij}(w)$ est le nombre d'occurrences de w dans le sous-segment S_{ij}. Cette technique de calcul de la probabilité généralisée peut éventuellement être couplée avec une technique d'interpolation lors du calcul du modèle de langue Δ_{ij} comme nous l'avons présenté dans la section 5.1.3.

7.2.1.1 Normalisation

La première piste mise en œuvre pour gérer la distribution des mots dans le document à segmenter consiste à normaliser la probabilité d'apparition d'un mot w dans le sous-segment S_{ij} par la probabilité que ce mot soit dans le segment S_i. Notre objectif ici est de privilégier les mots apparaissant localement et de diminuer le poids des mots présents un peu partout dans le segment S_i. Le calcul de la probabilité généralisée est modifié comme suit :

$$\ln P[S_{ij};\Delta_{ij};\Delta_i] = \sum_{w \in S_i} n_{ij}(w) \ln \frac{P[w;\Delta_{ij}]}{P[w;\Delta_i]} \ , \tag{7.3}$$

avec $P[w;\Delta_{ij}]$ (resp. $P[w;\Delta_i]$) la probabilité d'apparition du mot w dans le segment S_{ij} (resp. S_i).

Le tableau 7.2 présente les valeurs de la probabilité d'apparition de deux mots u et v dans le segment S_{ij} en fonction de leur nombre d'occurrences dans le segment S_i et dans le sous-segment S_{ij}. Ce tableau nous montre que les valeurs de cette probabilité sont égalee pour les mots u et v lorsque la probabilité généralisée est calculée grâce à l'équation 7.2, et ne traduit donc pas le fait que le mot u n'apparaisse qu'au sein du segment S_{ij} et qu'il est donc potentiellement caractéristique du contenu de ce sous-segment. La normalisation de la probabilité, présentée dans l'équation 7.3, fournit les valeurs de la colonne de droite du tableau. Dans cette colonne, nous constatons que le mot u, dont toutes les occurrences se situent au sein du sous-segment S_{ij}, est associé à une valeur de probabilité généralisée plus importante que le mot v, possédant le même nombre d'occurrences dans S_{ij} mais ayant en plus deux occurrences dans le segment de niveau hiérarchique supérieur. La normalisation de la probabilité généralisée permet donc, en théorie, de donner plus de poids à un mot présent seulement dans le sous-segment par rapport à un terme ayant une fenêtre d'apparition plus large dans le document.

7.2.1.2 Divergence

Afin de pénaliser les mots qui apparaissent dans tout le segment S_i, et qui sont donc moins caractéristiques du sous-segment S_{ij}, nous proposons, dans un deuxième temps, de modifier le calcul de la probabilité généralisée en nous inspirant de la divergence de Kullback-Leibler (Kullback and Liebler, 1951). Cette divergence, qui mesure la dissimilarité entre deux distributions de probabilités, va nous permettre de comparer la distribution des probabilités d'apparition de w dans S_{ij} et dans S_i. S'il existe une grande dissimilarité entre ces deux distributions de probabilités, cela signifie que w est présent (presque) exclusivement dans S_{ij}. Cette modification a pour but de pénaliser les mots apparaissant fréquemment dans la totalité du segment thématique de niveau hiérarchique supérieur. De ce fait, elle joue un rôle assez similaire à celui de l'*inverse document frequency* utilisée en recherche d'information pour pondérer les termes d'indexation des documents. Formellement, la probabilité généralisée est redéfinie de la façon suivante :

$$\ln P[S_{ij};\Delta_{ij};\Delta_i] = \sum_{w\in S_i} P[w;\Delta_{ij}] \ln \frac{P[w;\Delta_{ij}]}{P[w;\Delta_i]} \ . \tag{7.4}$$

Cette nouvelle mesure de la probabilité généralisée ne prend cependant plus en compte le nombre d'occurrences du mot w dans le segment S_{ij} ; nous le réintroduisons dans l'équation suivante :

$$\ln P[S_{ij};\Delta_{ij};\Delta_i] = \sum_{w\in S_i} n_{ij}(w) P[w;\Delta_{ij}] \ln \frac{P[w;\Delta_{ij}]}{P[w;\Delta_i]} \ . \tag{7.5}$$

Finalement, afin de pénaliser davantage les mots non caractéristiques du sous-segment, nous proposons d'évaluer la différence entre la distribution de probabilités des mots dans le sous-segment S_{ij} et la distribution des mots dans la transcription de toute l'émission T. De cette manière, nous espérons accentuer les différences entre les distributions de probabilités et donner, ainsi, davantage de poids aux mots présents uniquement dans le sous-segment S_{ij}. Formellement, on a :

$$\ln P[S_{ij};\Delta_{ij};\Delta_i] = \sum_{w\in S_i} n_{ij}(w) P[w;\Delta_{ij}] \ln \frac{P[w;\Delta_{ij}]}{P[w;\Delta_T]} \ , \tag{7.6}$$

avec Δ_T le modèle de langue estimé sur la transcription T de toute l'émission de reportages considérée.

7.2.1.3 Proportion

Dans (Eisenstein, 2009), Eisenstein se base sur l'idée que si un mot w apparaît localement dans le document entre les positions p_1 et p_2, alors le modèle de langue appris sur cette exacte portion de texte est plus à même de représenter w qu'un modèle de langue appris sur une portion du texte plus étendue. Afin de prendre en compte cette hypothèse, nous avons redéfini la probabilité généralisée de la façon suivante :

$$\ln P[S_{ij};\Delta_{ij};\Delta_i] = \sum_{w\in S_i} n_{ij}(w)(p(w)\ln P[w;\Delta_{ij}] + (1-p(w))\ln P[w;\Delta_i]) \ , \tag{7.7}$$

avec $p(w) = \frac{n_{ij}(w)}{n_i(w)}$ la proportion d'occurrences du mot w dans le sous-segment par rapport à son nombre d'occurrences dans le segment S_i. L'idée sous-jacente à cette définition est qu'un

TAB. 7.3 – Comparaison des différents niveaux de granularité dans le corpus d'*Envoyé Spécial*.

	nombre de frontières	durée moyenne des segments	nombre moyen de mots pleins répétés par segments	nombre de mots pleins dans chaque segment
hiérarchie supérieure	26	32 min *max* : 55 min, *min* : 22 min	140	1639
hiérarchie inférieure	246	3,4 min *max* : 20 min, *min* : 7 sec	15	173

mot est plus ou moins bien représenté par le modèle de langue du sous-segment S_{ij}, en fonction de sa proportion d'apparitions dans ce sous-segment par rapport au reste du segment. Cette modification de la probabilité généralisée va permettre de donner plus de poids au modèle de langue appris sur le seul sous-segment S_{ij} pour les mots apparaissant davantage dans ce sous-segment qu'ailleurs dans S_i, c'est-à-dire ceux pour lesquels la valeur de p est proche de 1.

Afin de mettre en avant les particularités du sous-segment S_{ij}, cette nouvelle définition de la probabilité généralisée peut également prendre en compte la proportion d'apparitions du mot w dans le segment S_i par rapport à son nombre d'occurrences dans la transcription complète T, en remplaçant $p(w)$ par $p'(w) = \frac{n_{ij}(w)}{n_T(w)}$ et le modèle de langue Δ_i par Δ_T dans l'équation précédente.

7.2.1.4 Résultats

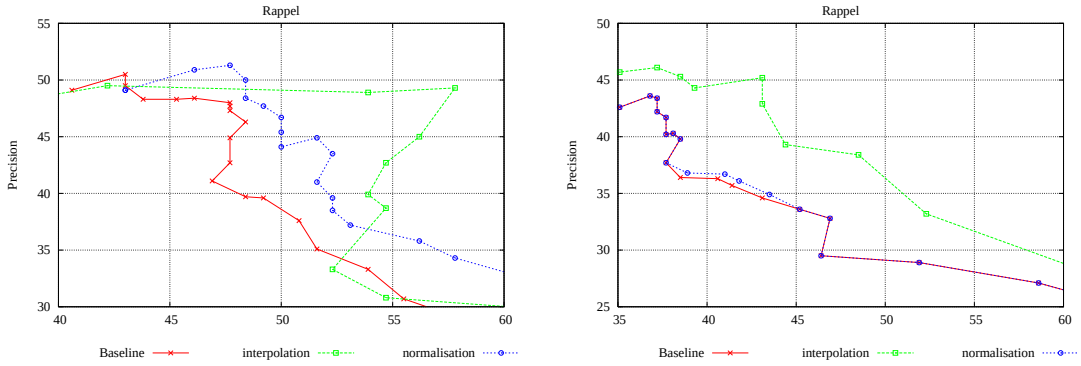
Les trois modifications de la probabilité généralisée décrites précédemment ont été testées sur sept émissions *Envoyé Spécial* transcrites automatiquement par le système IRENE. Ce corpus a été préféré à ceux précédemment utilisés car il présente une structure hiérarchique plus importante, chaque reportage étant découpé en différents points de vue, ce qui n'est pas toujours le cas de ceux présents dans les journaux télévisés et les émissions *Sept à Huit*. Les segmentations de référence ont été obtenues en découpant manuellement les émissions en thèmes et sous-thèmes, donnant lieu à deux niveaux de hiérarchie, le premier contenant 26 frontières et le second 246[5]. Si les segments du niveau de hiérarchie supérieur sont longs et de taille relativement stable, ceux du second niveau possèdent des caractéristiques plus proches de celles des corpora *JT* et *Sept à Huit*, c'est-à-dire segments courts, peu de répétitions, *etc.* comme le montre le tableau 7.3.

Les différentes modifications ont également été testées sur un sous-ensemble de quatre émissions transcrites manuellement. L'utilisation de ce corpus de transcriptions manuelles va nous permettre d'isoler les effets liés aux erreurs de transcription de l'analyse des résultats fournis par les changements dans la méthode de calcul de la probabilité généralisée.

Une segmentation de niveau hiérarchique supérieure à été obtenue pour les émissions, transcrites manuellement et automatiquement, grâce à l'algorithme de Utiyama et Isahara. Les performances obtenues pour cette segmentation, rappel et précision égaux à 100% dans le cas des transcriptions manuelles et à 94,7%[6] pour les transcriptions automatiques, nous

[5]Nous tenons à remercier Monica Corlay pour le travail effectué sur l'annotation et la segmentation thématique manuelle des différentes émissions.

[6]Dans le cas des transcriptions automatique, une seule erreur à été détectée dans la segmentation de niveau



transcriptions manuelles transcriptions automatiques

FIG. 7.1 – Segmentation thématique de segments thématiquement homogènes utilisant la méthode classique, une interpolation des modèles de langues et une normalisation de la probabilité généralisée

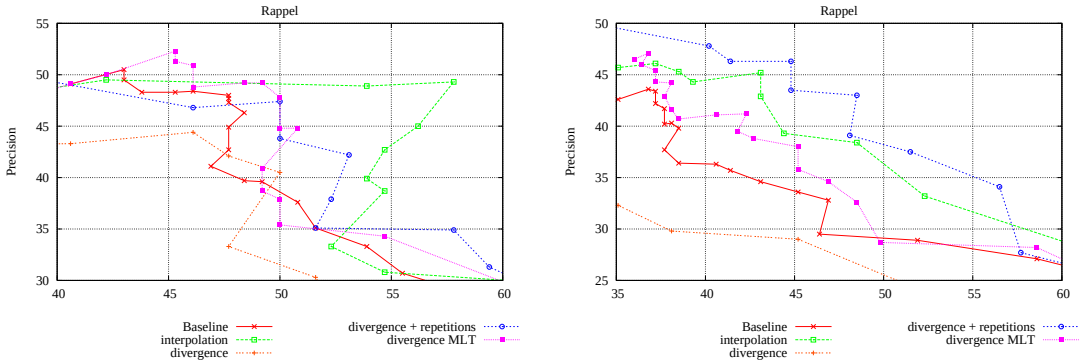
permettent de considérer que la qualité de la segmentation de niveau hiérarchique supérieure n'a pas d'influence, dans ce travail, sur celle des segmentations de niveaux inférieurs.

Nous avons tout d'abord évalué l'influence de la normalisation de la probabilité généralisée, que nous avons comparée avec l'impact de l'interpolation des modèles de langue présentée au chapitre 5. L'interpolation des modèles de langues (associée à un poids $\lambda = 0,4$), qui a montré son utilité pour gérer une taille de segments thématiques faible, permet d'améliorer la qualité de la segmentation des segments de niveau hiérarchique supérieur en sous-segments thématiques. En effet, la figure 7.1 nous montre que, pour les transcriptions manuelles et automatiques, la courbe verte représentant l'interpolation est bien au-dessus de la rouge correspondant à la segmentation obtenue avec un calcul de la cohésion lexicale classique. Sur cette figure, nous pouvons également constater que les résultats obtenus pour la segmentation en sous-segments sont relativement faibles, la valeur de la mesure F_1 étant égale à 37,7 sur les transcriptions automatiques et à 47,7 sur les transcriptions manuelles. Cette faible qualité des résultats justifie la mise en place de méthodes pour adapter l'algorithme de segmentation thématique linéaire à la tâche de segmentation thématique hiérarchique.

Cette figure nous présente également les résultats fournis par la normalisation de la probabilité généralisée (courbe bleue). Nous constatons que l'impact de cette modification est beaucoup plus élevé sur les transcriptions manuelles que sur les transcriptions automatiques, la normalisation n'ayant presque pas d'influence sur ces dernières. Cette différence est liée aux erreurs de transcription qui ont un impact négatif important sur la normalisation de la probabilité généralisée. En effet, un mot w apparaissant une fois dans le sous-segment S_{ij} et deux fois dans le segment S_i est associé à une valeur de probabilité généralisée normalisée égale à 0,52 et n'est pas considéré comme caractéristique du sous-segment, la moitié de ses occurrences étant située à l'extérieur du sous-segment. Cependant, si l'occurrence située dans le segment S_i n'est pas reconnue par le système de reconnaissance de la parole, le mot w va être considéré comme étant très caractéristique du segment – et associé à une valeur de probabilité généralisée de 1,22 – toutes ses occurrences apparaissant à l'intérieur du sous-segment.

Nous avons également étudié l'impact de la modification inspirée de la divergence de

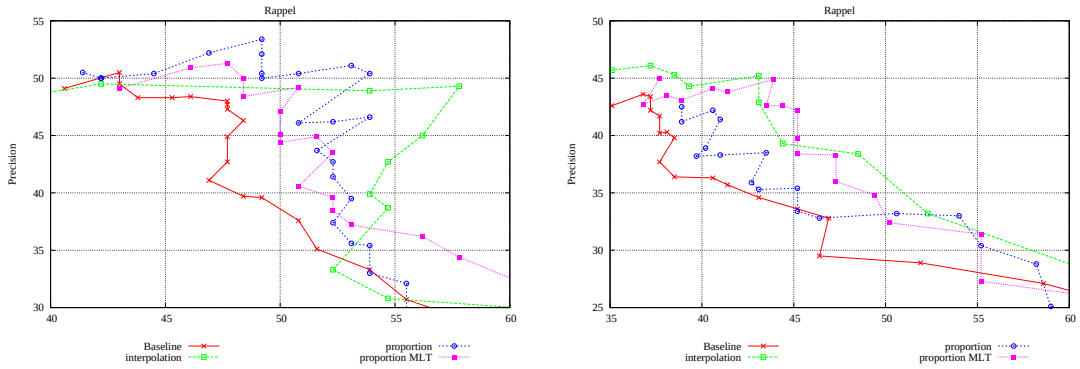
hiérarchique supérieur : l'une des frontières dans l'émission diffusée de 10 octobre 2008 est décalée de 5 minutes par rapport à la segmentation de référence.



transcriptions manuelles transcriptions automatiques

FIG. 7.2 – Segmentation thématique de segments thématiquement homogènes utilisant la méthode classique et une modification de la probabilité généralisée inspirée de la divergence de Kullback-Liebler

Kullback-Liebler sur les corpora transcrits manuellement et automatiquement. Les résultats obtenus grâce à cette modification, présentés sur la figure 7.2, montrent que la divergence, telle qu'elle est définie dans l'équation (7.4), dégrade les résultats de la segmentation (courbe verte). Cette dégradation est liée au fait que la définition ne prend pas en compte le nombre d'occurrences des mots dans les segments ; or la cohésion lexicale calculée dépend fortement de la répétition des mots. En ré-introduisant cette information (équation (7.5)), nous constatons que la modification de la probabilité généralisée améliore de façon significative les performances de l'algorithme de segmentation. Cette amélioration est par ailleurs plus marquée sur les transcriptions automatiques – et surpasse les résultats obtenus grâce à la technique d'interpolation des modèles de langue – que manuelles ce qui nous laisse à penser que les distributions de probabilités au sein des transcriptions manuelles sont plus similaires dans ces données. Cette observation n'est pas surprenante si l'on considère que les erreurs de transcription introduisent dans les transcriptions automatiques des mots totalement indépendants du reste du vocabulaire utilisé dans le document, alors qu'il existe une certaine cohérence dans le vocabulaire des segments transcrits manuellement. La dissimilarité entre les distributions de probabilité a également été calculée en comparant les probabilités d'apparition des mots dans le sous-segment S_{ij} sachant le modèle de langue Δ_{ij} et la probabilité de leur apparition sachant le modèle de langue Δ_T estimé sur l'émission complète. Cette dernière expérience fournit des résultats moins satisfaisants pour les deux types de transcription. Les courbes notées *MLT* (pour Modèle de Langue estimés sur la Transcription complète) sont, en effet, en-dessous des courbes *divergence + repetitions*. Cette tendance s'explique, selon nous, par le fait que, lors de la comparaison avec la transcription de l'émission complète, trop de mots sont pénalisés. En effet, comme nous l'avons montré dans le tableau 7.1, certains mots apparaissent exclusivement dans les sous-segments qu'ils caractérisent, comme *hassan* ou *légume*. Cependant, des mots comme *reportage* ou *soir*, qui sont effectivement caractéristiques de leur sous-segments et n'apparaissent (presque) pas ailleurs dans le segment englobant, sont présents dans le reste de la transcription. Ainsi, 80% des occurrences du mot *soir* apparaissant dans le segment sont situées dans le sous-segment mais seulement 25% de celles trouvées dans l'émission entière sont localisées dans le sous-segment. Le poids de ces mots dans le calcul de la cohésion lexicale va donc être diminué de façon inappropriée.



transcriptions manuelles transcriptions automatiques

FIG. 7.3 – Segmentation thématique de segments thématiquement homogènes utilisant la méthode classique et une modification de la probabilité généralisée prenant en compte la proportion d'apparitions des mots à l'intérieur et à l'extérieur des sous-segments

Finalement, nous avons adapté l'algorithme de segmentation thématique linéaire à une tâche de segmentation hiérarchique à partir de l'idée qu'un mot est plus ou moins bien représenté par le modèle de langue du sous-segment S_{ij}, en fonction de sa proportion d'apparitions dans ce sous-segment par rapport au reste du segment. À partir des résultats présentés sur la figure 7.3, nous pouvons constater que la méthode *proportion* permet d'améliorer la qualité de la segmentation pour les deux types de transcription. Pour les transcriptions manuelles, nous constatons que les meilleurs résultats sont obtenus lorsque le facteur $p(w)$ représente la proportion d'apparitions du mot w dans le sous-segment par rapport au segment et non par rapport à l'émission complète, ce qui s'explique par le fait que ce calcul de proportion d'apparition de w dans le sous-segment a plus de sens, comme nous l'avons expliqué dans le paragraphe précédent. Cependant, nous observons la tendance inverse sur les transcriptions automatiques. En effet, dans ce cas, la cohésion lexicale calculée en prenant en compte la proportion d'apparitions des mots dans la transcription complète (courbe *proportion MLT*) fournit les meilleurs performances. Ce comportement peut être lié à la présence d'erreurs de transcription dans les transcriptions automatiques, ces erreurs dégradant l'estimation des modèles de langue Δ_i et Δ_T. Cependant, la quantité de transcriptions utilisées pour estimer le modèle de langue Δ_T étant plus importante que celle prise en compte lors du calcul de Δ_i, les erreurs de transcription sont lissées dans le premier cas. Ceci explique, selon nous, le fait que la technique prenant en compte la proportion d'apparitions des mots dans S_{ij} par rapport à T conduise à une plus grande amélioration des performances de l'algorithme.

Ainsi, la pénalisation des mots apparaissant tout au long du segment thématique de niveau hiérarchique supérieur permet de favoriser la prise en compte de mots caractéristiques des sous-segments lors du calcul de la cohésion lexicale. En effet, les modifications de la méthode du calcul de la probabilité généralisée augmentent les valeurs de la mesure F_1, pour les transcriptions automatiques, de 6,4 points dans le cas de la technique appelée *divergence* et de 5.3 points pour la *proportion*.

7.2.2 Chaînes lexicales

Dans la première partie de ce travail, nous avons développé différentes techniques visant à pénaliser les mots présents à la fois dans un sous-segment S_{ij} et dans le segment S_i, c'est-à-dire

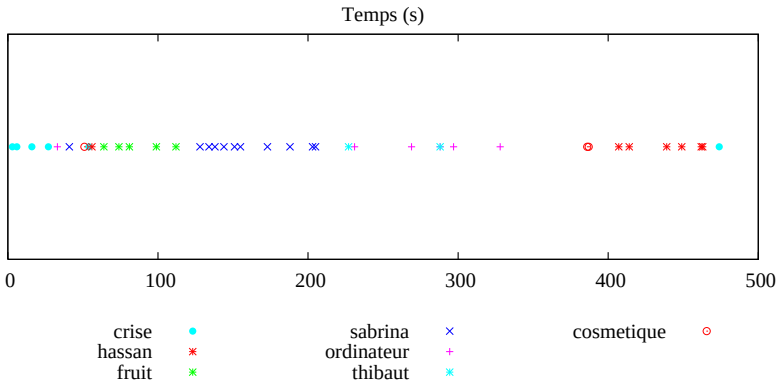


FIG. 7.4 – Position, dans la transcription automatique de « La France de la débrouille », de certains mots, très représentatifs des sous-thèmes abordés dans le reportage

ceux ayant déjà participé fortement au calcul de la cohésion lexicale du niveau hiérarchique supérieur. Or il est possible qu'un mot fortement caractéristique d'un sous-segment apparaisse à plusieurs endroit du document. La figure 7.4 présente les positions de certains mots extraits du premier reportage de l'émissions *Envoyé Spécial* du 15 janvier 2009. Ce reportage, traitant des conséquences de la crise économique en France, décrit successivement les histoires de plusieurs personnages : Toufik, Sabrina, Thibaut et Hassan. On constate sur cette figure que certains mots apparaissent de façon très localisée. Ainsi, les occurrences de « Sabrina » sont majoritairement présentes entre la 100^e et la 200^e position. De même, les occurrences des mots « Hassan » et « fruit » sont regroupées dans le document. Cependant, on constate que les mots « Sabrina » et « Hassan » apparaissent également au début du document, c'est-à-dire dans l'introduction du reportage. Notre objectif ici n'est donc pas de pénaliser les mots présents à plusieurs endroits du document mais plutôt de favoriser ceux qui apparaissent de façon localisée à un ou plusieurs endroits du document. Afin de prendre en compte cette localisation des mots, nous nous basons sur le principe des chaînes lexicales, comme proposé dans (Angheluta et al., 2002; Sitbon and Bellot, 2005) ou (Stokes et al., 2002).

7.2.2.1 Calcul des chaînes lexicales

Le calcul des chaînes lexicales s'effectue de la façon suivante : une chaîne lexicale est créée entre deux groupes de souffle b_1 et b_2 si un même mot (ou deux mots sémantiquement liés) apparaît dans b_1 et b_2, et si b_1 et b_2 sont séparés par moins de Γ secondes. À partir de ces chaînes lexicales, chaque frontière potentielle existant dans notre fichier à segmenter (c'est-à-dire chaque séparation entre deux groupes de souffle) est associée à une valeur qui prend en compte le nombre de chaînes lexicales qui « enjambent » cette frontière potentielle. Grâce à ces valeurs, nous souhaitons maximiser la cohésion des segments qui contiennent un nombre important de mots apparaissant de façon très localisée. La figure 7.5 montre que le nombre de chaînes lexicales associé à chaque frontière potentielle du reportage « La France de la débrouille » semble être corrélé avec la présence ou non d'une frontière thématique. Nous pouvons, en effet, constater qu'une frontière thématique, représentée par une ligne verticale rouge, est toujours associée à un faible nombre de chaînes lexicales (courbe bleue), les valeurs

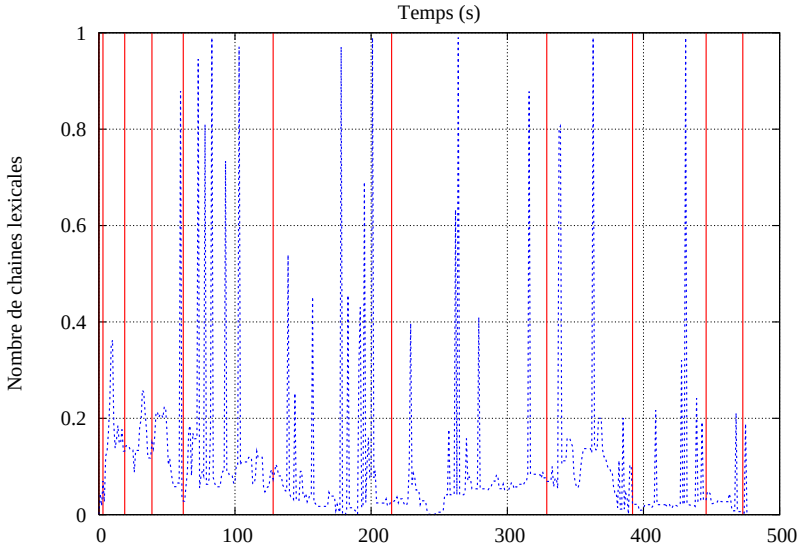


FIG. 7.5 – Nombre, normalisé entre 0 et 1, de chaînes lexicales, calculées à partir de la répétition de lemmes, qui « enjambent » chaque frontière potentielle dans le reportage « La France de la débrouille »

importantes étant toujours situées à l'intérieur d'un segment thématique.

Dans un premier temps, les chaînes lexicales ont été calculées seulement à partir des répétitions des lemmes. Puis, afin de prendre en considération les liens sémantiques qui peuvent exister entre deux termes différents, un ensemble de relations paradigmatiques extraites automatiquement et sélectionnées grâce à la méthode $ParMot_2$ (2 relations sémantiques par mot, *cf.* chapitre 3) a également été utilisé. Finalement, nous avons calculé les chaînes lexicales en nous basant sur la répétition des *stems*[7] plutôt que sur celle des lemmes pour prendre en considération les variations morphologiques tant flexionnelles que dérivationnelles.

7.2.2.2 Prise en compte des chaînes lexicales pour segmenter un segment thématiquement homogène

Comme dans le chapitre 4, nous exploitons le formalisme de l'algorithme de Utiyama et Isahara pour prendre en considération le nombre de chaînes lexicales associé à chaque frontière potentielle lors du calcul de la segmentation thématique hiérarchique. Dans ce cadre, l'objectif de la segmentation, décrit par :

$$\hat{S} = \underset{S}{\operatorname{argmax}}\, P[W|S]\ P[S]\ \ , \tag{7.8}$$

est redéfini de la façon suivante :

$$\hat{S} = \underset{S}{\operatorname{argmax}}\, P[W|S]\ P[S|L]^{\omega}\ P[S]\ \ . \tag{7.9}$$

avec $P[S|L]$ la probabilité d'obtenir une segmentation S connaissant les chaînes lexicales associées au document à segmenter. La valeur du coût d'un segment S_i constitué des groupes

[7]Les *stems* sont extraits grâce à Lingua : :Stem.

de souffle $s_a ... s_b$ devient donc

$$\begin{aligned} C(S_i|W,L) &= -logP[W_i|S_i] \\ &\quad -\omega \left[\sum_{j=a}^{b-1} log\ P\ (B_j = \text{« non »}|L_j) + log\ P\ (B_b = \text{« oui »}|L_b) \right] \\ &\quad -\alpha\ log\ P(S_i)\ , \end{aligned} \tag{7.10}$$

avec $P\ (B_j = \text{« non »}|L_j)$ (resp. $P\ (B_b = \text{« oui »}|L_b)$) la probabilité qu'il n'y ait pas de (resp. qu'il y ait une) frontière thématique après le groupe de souffle j (resp. b) sachant le nombre de chaînes lexicales « enjambant » ce groupe de souffle.

La première étape nécessaire à l'intégration de l'information relative au nombre de chaînes lexicales associées à chaque frontière potentielle consiste donc à calculer, pour chaque document à segmenter les log-probabilités qu'il existe une frontière thématique entre chacun des groupes de souffle. Ces log-probabilités sont évaluées de la façon suivante :

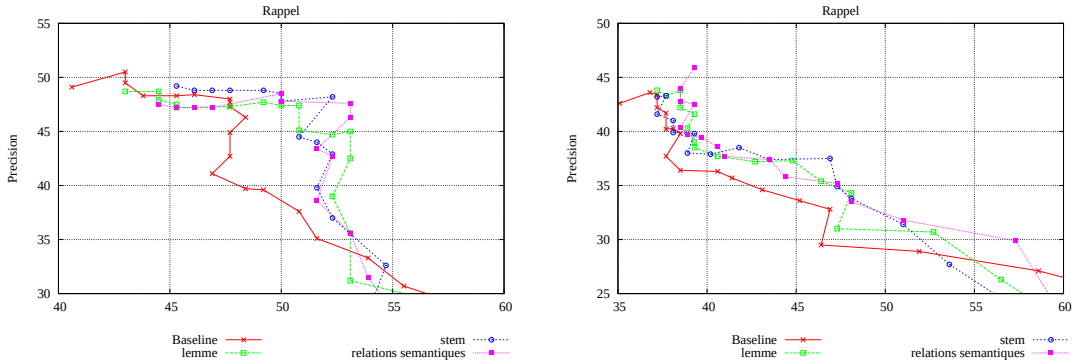
$$log\ P\ (B_b = \text{« oui »}|L_b) = log\ \left(1 - \frac{\sum_{cl} C(cl,b) - min_{100}}{max_{100} - min_{100}}\right)\ , \tag{7.11}$$

avec $C(cl,b)$ le poids de la chaîne lexicale cl pour la frontière potentielle b. Ce poids est égal au *ratio* entre la durée de cl en secondes et le nombre d'occurrences de mots qu'elle contient lorsque cl « enjambe » la frontière potentielle b, et à 0 sinon. max_{100} et min_{100} correspondent aux valeurs des poids maximaux et minimaux associés à une frontière potentielle située dans une fenêtre de 100 frontières potentielles centrée sur b.

Les chaînes lexicales introduites ont été calculées avec un hiatus Γ variant de 50 à 200 secondes et ont été employées, comme pour les modifications présentées dans la section précédente, pour la segmentation d'émissions *Envoyé Spécial* transcrites manuellement et automatiquement. Pour les trois types de chaînes lexicales utilisées, basées sur la répétition de lemmes, de *stems* ou prenant en compte des relations sémantiques, nous avons pu constater que la valeur optimale du paramètre Γ était égale à 150. C'est, en effet, pour cette valeur que les améliorations constatées sont les plus importantes (*cf.* annexe D). La figure 7.6 résume les performances de l'algorithme de segmentation appliqué sur des segments thématiquement homogènes et intégrant les différents types de chaînes lexicales avec cette valeur de hiatus. Sur cette figure, nous remarquons que, si les chaînes lexicales permettent d'améliorer la qualité de la segmentation thématique de façon statistiquement significative, il n'existe pas de réelle différence entre les trois types de chaînes lexicales utilisées, que ce soit pour les transcriptions manuelles ou les transcriptions automatiques. Les courbes représentant l'intégration des chaînes lexicales sont, en effet, quasiment confondues. Si l'amélioration proposée par les chaînes lexicales est un peu plus faible pour les transcriptions automatiques que pour les transcriptions manuelles, l'influence des chaînes lexicales sur les deux corpora est tout à fait comparable. Nous pouvons donc conclure de ces remarques que la prise en compte de la fenêtre d'apparition des mots dans un segment thématique est un indice utile pour améliorer les performances d'un algorithme de segmentation thématique linéaire dans le cadre d'une tâche de segmentation thématique hiérarchique.

7.3 Perspectives

Le travail proposé dans ce chapitre étant exploratoire, de nombreuses pistes restent à étudier pour adapter notre algorithme de segmentation thématique linéaire à une tâche de



transcriptions manuelles transcriptions automatiques

FIG. 7.6 – Intégration des chaines lexicales

segmentation thématique hiérarchique.

Tout d'abord, la prise en compte de la distribution des termes au sein des documents à segmenter fournissant des résultats plutôt encourageants, il nous semble intéressant de poursuivre plus avant les expérimentations mises en place. Pour ce faire, l'utilisation de techniques de calcul de chaînes lexicales plus sophistiquées peut être envisagée. En effet, si la prise en compte des seules répétitions de lemmes ou de *stems* permet d'augmenter les valeurs de la mesure F_1 de 1,4 de points et de 1,8 de points respectivement, ces méthodes ne gèrent pas la résolution d'anaphores pronominales. La longueur d'une chaîne lexicale ou son poids, qui dépend du nombre d'occurrences du mot qu'elle contient, peut ainsi être diminué faute de prise en compte des liens existant entre un nom et son pronom référent[8]. Une modification des poids associés à ces chaînes doit également être envisagée. En effet, une chaîne reliant les occurrences d'un mot apparaissant tout au long du document à segmenter devrait être associée à un poids relativement faible, le mot considéré étant *a priori* plus caractéristique du thème que des sous-thèmes. Cette modification du poids des chaînes lexicales prenant en compte l'amplitude permettrait ainsi de combiner les deux types de techniques proposées dans ce chapitre.

De plus, une perspective à plus long terme consiste à mettre en place un algorithme capable de fournir en une seule passe une segmentation possédant plusieurs niveaux de hiérarchie, dont le nombre serait, éventuellement, fourni par l'utilisateur. En effet, si l'algorithme développé par Eisenstein (Eisenstein, 2009) propose une segmentation à plusieurs niveaux de hiérarchie, son fonctionnement ne semble pas adapté à nos données, la taille des segments retournés étant trop régulière. La mise en place d'un tel algorithme peut passer par l'adaptation de l'algorithme de Utiyama et Isahara afin de calculer une segmentation globale qui maximiserait à la fois la cohésion au sein des segments thématiques et au sein des sous-segments. Cependant, un tel algorithme nécessite de rechercher le meilleur chemin au sein d'un graphe valué représentant

[8]Lors d'une étude préliminaire sur la prise en compte des pronoms lors de la segmentation thématique d'émissions télévisuelles nous avons constaté que, malgré les erreurs de transcriptions présentes dans nos données, les positions des pronoms au sein des documents pouvait constituer une aide. En effet, les pronoms n'apparaissaient qu'au milieu ou à la fin des segments thématiques, jamais en début de segment. Bien que des expériences préliminaires consistant à pénaliser les frontières thématiques placées avant un groupe de souffle contenant plusieurs pronoms n'ont pas conduit à des améliorations importantes de la qualité de la segmentation, nous pensons néanmoins que ces informations peuvent être utiles pour diminuer le nombre de frontières mal positionnées.

toutes les segmentations possibles ainsi que toutes les sous-segmentations associées à chacune de ces segmentations. Face à cette difficulté, une possibilité consiste à s'inspirer de techniques développées dans des domaines connexes au nôtre, comme dans (Slaney and Ponceleon, 2001) où les auteurs s'inspirent de la segmentation de signal audio pour développer une méthode de segmentation thématique hiérarchique. La technique de segmentation d'image dite *ligne de partage des eaux*, qui consiste à représenter l'image à segmenter comme un relief et à identifier dans ce dernier la ligne de partage des eaux[9], ayant montré son intérêt dans le cadre de la segmentation thématique linéaire (Claveau and Lefèvre, 2011b), il pourrait être intéressant de chercher à adapter cette technique à la problématique de la segmentation thématique hiérarchique en prenant en compte des variations plus subtiles dans le relief.

L'un des problèmes que pose le travail de la segmentation thématique hiérarchique est son évaluation. Dans ce chapitre, nous avons utilisé une technique d'évaluation de segmentation linéaire ne permettant pas de comparer deux segmentations hiérarchiques. En effet, les évaluations fournies comparent seulement le second niveau de hiérarchie, la segmentation de niveau hiérarchique supérieur étant fixée *a priori*. Afin d'évaluer globalement les résultats d'un algorithme de segmentation hiérarchique, la technique d'évaluation pourrait s'inspirer de celle développée pour des travaux d'appariement de documents XML. Ces travaux consistent à calculer la distance d'édition minimum permettant de transformer la structure d'un document XML F en F', F' étant parfaitement défini par un schéma XML. En représentant la segmentation hiérarchique de référence sous la forme d'un document XML F, l'évaluation de la segmentation thématique hiérarchique pourrait être vue comme le calcul de la distance d'édition entre F et une segmentation hiérarchique hypothèse représentée sous la forme d'un document XML F'. De cette manière, la similarité entre une segmentation hypothèse et la segmentation de référence prendrait en compte l'aspect structurel de la segmentation thématique hiérarchique.

Finalement, la structuration obtenue grâce à la segmentation thématique hiérarchique doit pouvoir être exploitée pour mettre en relation des sous-segments thématiquement homogènes comme nous l'avons vu dans le chapitre précédent. Cependant, la mise en relation de segments de plus en plus petits, et contenant de ce fait de moins en moins de vocabulaire, ne pourra pas être abordé uniquement par le biais du seul vocabulaire présent dans le sous-segment. Afin de pallier cette difficulté, une des possibilités existante consiste à caractériser le sous-segment par des vecteurs composés non seulement des mots contenu dans le sous-segment mais également de ceux apparaissant dans le segment englobant, le poids des mots dans le vecteur pouvant être modifier en fonction de la proportion d'apparitions du mot dans le sous-segment par rapport au segment.

7.4 Bilan du chapitre

Dans ce chapitre, nous avons proposé deux types de modifications du calcul de la cohésion lexicale pour adapter l'algorithme de segmentation thématique linéaire de Utiyama et Isahara à une tâche de *structuration thématique hiérarchique*. La première approche consiste à favoriser les mots caractéristiques des sous-segments, c'est-à-dire ceux apparaissant principalement dans le sous-segment, en pénalisant les mots présents dans tout le segment thématique. La seconde technique prend en compte la distribution du vocabulaire au sein des segments thématiques à segmenter grâce à l'utilisation de chaînes lexicales. Ces deux méthodes ont conduit à une

[9]La ligne de partage des eaux est la limite qui divise un territoire en un ou plusieurs bassins versants.

amélioration de la qualité de la segmentation hiérarchique, avec une augmentation de la valeur de la mesure F_1 de 6,4 points pour la modification du calcul de la probabilité généralisée et de 1,8 points pour l'utilisation des chaînes lexicales. Le travail exploratoire mené dans ce chapitre a ainsi permis de montrer, de façon quantitative, qu'un algorithme de segmentation thématique linéaire basé sur la cohésion lexicale pouvait être employé dans le cadre de la segmentation hiérarchique moyennant une modification du calcul de la cohésion lexicale qui permet d'une part de pénaliser les mots ayant participé au calcul de la cohésion lexicale lors de la segmentation à des niveaux hiérarchiques supérieurs et de favoriser les termes qui apparaissent localement au sein des segments thématiques à segmenter.

Conclusion

L'objectif de cette thèse a été d'étudier la mise en place de techniques de structuration automatique de flux multimédia. Plus particulièrement, le nombre de documents multimédia disponibles ne permettant pas la mise en place de méthodes *ad hoc* pour chaque type d'émission, nous nous sommes fixée comme contrainte de rester dans un cadre de travail le moins supervisé possible, en évitant toute connaissance *a priori* sur la structure des émissions traitées. Pour cela, nous nous sommes basée sur les transcriptions automatiques de la parole contenue dans les programmes télévisés sur lesquelles nous avons employé des méthodes de traitement automatique des langues adaptées à ces données particulières. Dans ce cadre, nous avons proposé deux types de structuration thématique, linéaire et hiérarchique, reposant sur une segmentation thématique linéaire des émissions télévisuelles. Ces deux objectifs de structuration, ainsi que la mise en place d'une segmentation thématique des programmes TV, ont fait l'objet de plusieurs contributions originales.

Tout d'abord, la phase de segmentation thématique linéaire constituant la base de nos techniques de structuration, plusieurs extensions de l'algorithme proposé par Utiyama et Isahara (Utiyama and Isahara, 2001) pour du texte écrit ont été mises en œuvre. Dans un premier temps, nous avons proposé de prendre en compte de façon conjointe deux indices généralement employés séparément dans les méthodes de segmentation thématique de l'état de l'art. La combinaison d'informations de rupture de la cohésion lexicale et de la mesure de cette cohésion a montré un gain dans les performances de la segmentation thématique. Cependant, cette contribution n'ayant pas pour objectif de gérer les spécificités des données télévisuelles transcrites, nous avons cherché dans un deuxième temps, à adapter la méthode de calcul de la cohésion lexicale aux particularités des transcriptions automatiques de vidéos professionnelles – segments thématiques potentiellement courts, faibles répétitions de vocabulaire et erreurs de transcription. Pour cela, nous avons intégré deux types d'informations complémentaires, les mesures de confiance fournies par le système de reconnaissance automatique de la parole et des relations sémantiques extraites de façon non supervisée. En comparant l'impact de ces indices sur des corpora contenant plus ou moins d'erreurs de transcription, nous avons pu montrer que les mesures de confiance permettaient effectivement de pénaliser les mots mal transcrits lors du calcul de la cohésion lexicale. De plus, nous avons mis en évidence le fait que les relations sémantiques avaient non seulement un impact sur la faible répétition de vocabulaire mais qu'elles permettaient, également, de pallier les erreurs de transcription. Une technique originale consistant à utiliser des méthodes d'interpolation des modèles de langue nous a également permis d'estimer de façon plus robuste la cohésion lexicale sur les segments thématiques de petite taille et d'améliorer, ainsi, la qualité de la segmentation de façon significative sur les émissions composées de segments thématiques très courts (quelques secondes). Finalement, notre dernière contribution concernant l'adaptation du calcul de la cohésion lexicale se fonde sur l'exploitation d'un autre aspect de la parole à travers la pro-

sodie. En intégrant des informations acoustiques tels que le *pitch* et l'*intensité*, nous avons pu améliorer les performances de l'algorithme en donnant plus de poids, lors du calcul de la cohésion lexicale, aux mots proéminents dans le discours.

À partir des segments thématiquement homogènes ainsi obtenus nous avons développé une *structuration thématique linéaire* consistant à mettre en relation des segments, extraits d'une collection de documents, abordant des sujets similaires. Cette technique de structuration, basée sur les transcriptions automatique de la parole contenue dans les segments, se décompose en deux étapes : la représentation des segments par des vecteurs de mots caractéristiques et le calcul de la similarité entre ces vecteurs. Dans ce cadre, nous avons proposé d'associer, à des méthodes de recherche d'information classiques, des indices propres aux documents oraux afin de faire émerger les termes prononcés avec plus d'emphase par le locuteur, généralement associés à une valeur informative importante. Les expériences menées ont démontré l'utilité de la prosodie pour la mise en relation de segments thématique en langue française. La technique de *structuration thématique linéaire* a également été adaptée à nos émissions télévisuelles par le biais de relations sémantiques. Ces relations, introduites originellement pour améliorer le calcul de la similarité entre les vecteurs abordant des sujets similaires par le biais d'un vocabulaire différent, ont surtout montré leur intérêt dans la gestion des erreurs de transcription présentes dans nos données. Une comparaison entre les résultats obtenus sur des transcriptions manuelles et automatiques a, en effet, mis en évidence le fait que les relations sémantiques extraites sans supervision permettaient de compenser les erreurs contenues dans des corpora affichant des taux d'erreurs de VALEUR et VALEUR2.

Finalement, dans une dernière partie plus exploratoire, nous avons étudié la question du passage de la segmentation thématique linéaire à la *structuration thématique hiérarchique*. Pour ce faire, nous avons utilisé de façon répétitive l'algorithme de segmentation thématique linéaire tout en adaptant le calcul de la cohésion lexicale afin de favoriser les mots caractéristiques des sous-segments lors de l'estimation de la cohésion. Dans ce contexte, nous avons proposé deux types de modifications de l'estimation de la cohésion lexicale. Notre première contribution a consisté à pénaliser les mots ayant déjà participé au calcul de la cohésion lexicale des segments thématiques des niveaux hiérarchiques supérieurs en modifiant la technique de calcul de la probabilité généralisée. Deuxièmement, nous avons utilisé des chaînes lexicales pour favoriser les mots apparaissant localement au sein des segments thématiques à segmenter afin de prendre en compte la distribution du vocabulaire. Ces différentes expériences ont menées à des améliorations de la qualité de la segmentation thématique hiérarchique, améliorations évaluées de façon quantitative sur deux corpora composés d'émissions de reportages transcrites manuellement et automatiquement.

Perspectives

Les pistes explorées dans cette thèse mettent au jour diverses perspectives de travail. Nous présentons dans cette dernière section celles qui nous semblent les plus pertinentes.

Tout d'abord, des améliorations pourraient être apportées aux différentes étapes nécessaires au développement des structurations thématiques, qu'elles soient linéaire ou hiérarchique. Dans un premier temps, les mesures de confiance ayant montré leur intérêt pour pénaliser les mots erronés lors de la segmentation thématique, les sorties intermédiaires du système de transcription nous semble être des indices importants à prendre en compte pour compenser les erreurs de transcription. En effet, il a été montré que, dans le cas d'une erreur

de transcription dans la sortie finale, le mot à reconnaître apparaissait fréquemment dans les premières hypothèses proposées par le système[10]. L'une des premières pistes envisagées consiste donc à fonder la segmentation thématique, non pas sur la seule transcription finale, mais sur la liste des n meilleures hypothèses proposées par le système de reconnaissance automatique de la parole. Dans un deuxième temps, un effort sur le calcul des relations sémantiques devrait être effectué. En effet, plusieurs des techniques mises en œuvre au cours de cette thèse utilisent des relations sémantiques apprises automatiquement. Or, les relations extraites n'étant pas toujours de très bonne qualité, nous avons dû développer des techniques de filtrage et de sélection afin de limiter le bruit introduit par des relations non pertinentes. La qualité des relations introduites semblant avoir un impact important sur la qualité des résultats obtenus, l'extraction de ces relations devrait s'appuyer sur des méthodes plus sophistiquées. Pour cela, nous pourrions, d'une part acquérir les relations sémantiques sur des corpora thématiquement homogènes plutôt que sur un corpus « de la langue générale » comme le propose (Pichon and Sébillot, 1999). D'autre part, l'impact du type de relations sémantiques introduites (synonyme, hyperonyme, *etc.*) reste à étudier. En effet, nous avons, dans nos travaux employé toutes sortes de relations paradigmatiques sans chercher à determiner si un type de relation était plus pertinent pour nos tâche de structuration. Finalement, la dernière partie de notre travail étant encore exploratoire, de nombreuses pistes restent à examiner concernant la segmentation thématique hiérarchique. Les améliorations observées lors des expériences menées sur la prise en compte de la distribution du vocabulaire au sein des documents à segmenter nous confortent dans l'idée qu'un tel indice doit être utilisé afin de définir un algorithme de segmentation thématique hiérarchique capable de produire, en une seule itération, une segmentation à plusieurs niveaux hiérarchiques.

Comme nous l'avons mentionné dans l'introduction de cette thèse, l'accès à l'information contenue dans un flux multimédia nécessite, d'une part, une extraction de l'organisation interne aux (collections de) documents considérés, et une représentation du contenu de ces documents d'autre part. Or, si la structuration a fait l'objet de la plupart des travaux présentés dans ce manuscrit, l'aspect description n'en reste pas moins essentiel. Dans le cadre des exemples d'applications proposées dans le chapitre 6, le contenu des journaux télévisés a été décrit grâce à des listes de mots clés extraits automatiquement des transcriptions par le critère *tf-idf*. Or ces mots clés, s'ils peuvent donner une idée du thème abordé dans le segment, entraînent parfois des erreurs de compréhension et offrent une description assez limitée du thème abordé dans le segment. Afin de proposer une représentation du contenu permettant aux utilisateurs d'accéder efficacement aux informations recherchées, une piste intéressante consiste à employer des méthodes de caractérisation plus sophistiquées que la simple liste de mots clés. Tout d'abord, il nous paraît important d'intégrer, dans la caractérisation du contenu des segments, les relations que les mots clés entretiennent les uns avec les autres. En effet, les mots *victoire*, *match*, *OM*, *PSG*, s'ils donnent une idée du contenu du segment, ne permettent pas de savoir rapidement qui du *PSG* ou de l'*OM* a remporté la victoire. Cette description peut passer par l'utilisation de techniques développées pour des tâches d'extraction de relations à partir de corpora textuels ou d'extraction d'information (Palmer et al., 2000). Cependant, ces méthodes, difficiles à mettre en œuvre, sont généralement très spécifiques à un domaine particulier et nécessitent souvent une phase d'apprentissage. Afin de dépasser une simple caractérisation sous forme de liste, il est également possible de fournir à l'utilisateur des

[10]Dans (Huet, 2007), Huet a en effet montré que le taux d'erreur de transcription obtenu sur les corpora ESTER 1 et ESTER 2 passait de 21,6 lorsqu'il était calculé sur les transcriptions finales à 13,9 lorsqu'il était évalué sur les graphes de mots.

résumés produits automatiquement. De nombreux travaux de résumé de documents audios ont développé des techniques qui pourraient être employées dans cet objectif.

Finalement, afin de permettre aux utilisateurs de naviguer efficacement au sein d'une collection de documents, il est nécessaire que la présentation des liens entre les documents soit suffisamment explicite : les documents abordent-ils le même thème, présentent-ils les évolutions d'un même sujet ou différents points de vue sur un même fait d'actualité ? Cette problématique de mise en place de moyens de navigation efficaces est un élément essentiel à la structuration automatique de flux multimédia, car sans moyen de compréhension de l'organisation obtenue toutes les méthodes de structuration développées deviennent inutiles. Pour mettre en évidence les différences existant entre deux segments abordant des thèmes similaires, certaines méthodes, développées dans le cadre du projet *Topic Detection and Tracking*, ayant étudié la détection de redondances et de nouveautés dans des flux d'actualités (Wu et al., 2006), pourraient être appliquées sur nos données. De même, les travaux de résumé multi-documents, comme le système *NewsBlaster* (McKeown et al., 2002) qui détecte à la fois les phrases communes aux différents segments traitant du même sujet et celles spécifiques à chacun des segments, peuvent être utilisés pour représenter le contenu commun à différents documents. Par ailleurs, la mise en place de moyens de navigation au sein d'une collection de documents thématiquement cohérents fait, à proprement parlé, l'objet d'un doctorat, en partenariat avec l'Institut National de l'Audiovisuel, commencé en 2011 par Ludivine Kuznik.

Annexe A

Définition du thème

La littérature sur le concept du thème et sa formalisation est abondante et conduit à un nombre très important de définitions. Face à cette multitude, nous choisissons de décrire plus précisèment deux de ces formalisations. Premièrement, nous nous intéressons au travail de Rastier, dont la théorie a souvent été utilisée en traitement automatique des langues, qui propose des définitions de la notion de thème traduisant la notion de granularité thématique que nous cherchons à prendre en compte dans le cadre de cette thèse. Deuxièmement, nous décrivons la théorie de Marandin (Marandin, 1988), fondée sur un travail lexicographique, qui propose à la fois une formalisation du thème mais également une méthode de description de ce thème par le bais d'un « mot-vedette », comme nous le présentons en A.2.

A.1 Définition de Rastier

Afin de définir la notion de thème, Rastier se base sur les concepts de la sémantique différentielle que nous rappelons brièvement ici.

Pour la sémantique différentielle, le sens d'un texte émerge d'une structuration de l'espace des sémèmes, qui sont les mots du lexique. Ces sémèmes sont définis les uns par rapport aux autres par des sèmes qui sont des relations sémantiques. Il existe deux catégories de sèmes, les sèmes génériques et les sèmes spécifiques. Pour les différencier, nous reprenons l'exemple proposé dans (Rossignol and Sébillot, 2003). Les sèmes génériques révèlent le partage par deux sémèmes d'un élément de sens ; ainsi, « autobus » et « autocar » se trouvent rapprochés par le sème /moyen de transport/. L'ensemble des sémèmes liés par un même sème générique constitue une classe sémantique. Les sèmes spécifiques marquent, pour leur part, les différences sémantiques existant entre sémèmes d'une même classe sémantique. Dans l'exemple précédent, « autobus » se distingue d'« autocar » par le sème spécifique /extra-urbain/.

Grâce à ces concepts, Rastier peut définir le thème de la façon suivante : le thème est « une structure stable de traits sémantiques (ou sèmes), récurrente dans un corpus, et susceptible de lexicalisations diverses ». La récurrence d'un sème générique induit une isotopie générique, qui est un effet de la récurrence syntagmatique d'un même sème. Selon Rastier, lorsque le mot « thème » est utilisé pour désigner le « sujet » d'un texte, il désigne en réalité son isotopie générique dominante. De plus, Rastier définit le thème spécifique comme une molécule sémique, c'est-à-dire un groupement de sèmes spécifiques. Cette différenciation entre thème spécifique et thème générique peut être vue comme une représentation des différentes granularité dans laquelle le thème spécifique serait un sous-thème du thème générique.

Le thème défini comme molécule sémique peut recevoir des lexicalisations diverses, qui peuvent aller du morphème au syntagme. Dans la voie sémantique, on va chercher d'autres mots et expressions qui sont cooccurrents et, une fois interprétés, les cooccurrents pour lesquels on aura identifié une relation sémantique seront considérés comme des corrélats, c'est-à-dire comme des lexicalisations complémentaires de la même molécule sémique. Le réseau de corrélats relie les manifestations lexicales du thème. Mais il faut pouvoir discerner le ou les meilleurs points d'entrée dans ce réseau. Ce (ou ces) point(s) d'entrée, qui permet(tent) de lexicaliser le thème, est/sont choisi(s) pour sa/leur fréquence.

L'étude statistique des cooccurrences lexicales pour l'analyse thématique se justifie selon Rastier par le fait que l'activation d'un trait sémantique favorise sa réitération, ce qui permet de constituer des isotopies génériques ou spécifiques. De la même manière, cette activation favorise la réitération des traits voisins de la même molécule sémique : c'est pourquoi les lexicalisations d'un même thème sont fréquemment cooccurrentes. Cela dit, bien que Rastier parte des mots pour définir un thème, il est selon lui difficile de passer de l'analyse lexicale à l'analyse thématique car une analyse thématique qui en resterait au palier lexical compterait potentiellement autant de thèmes que de mots de la langue. Or un mot du lexique peut ne lexicaliser aucun thème dans un corpus donné, mais il peut également en lexicaliser plusieurs.

A.2 Définition de Marandin

Dans son article (Marandin, 1988), Marandin étudie le texte à deux niveaux afin de proposer sa représentation du thème. Tout d'abord, il met en place une approche de morphologie discursive à travers l'étude de la chaîne-objet, puis il travaille au niveau de la compréhension des énoncés en différenciant thème configuré et thème inféré. Ces trois notions – chaîne-objet, thème inféré et thème configuré – vont être explicitées par la suite.

Marandin étudie la manière dont le thème d'un texte peut être compris grâce au concept de chaîne-objet emprunté à Chastain (Chastain, 1975). Une chaîne-objet est définie comme un dispositif de constitution de la dénotation des termes, c'est-à-dire comme le « contenu descriptif (CD) d'un terme (singulier) relativement à un discours donné ». Le contenu descriptif d'un terme singulier est ce que le discours dit de son référent, quelles propriétés il est supposé avoir, *etc.* Marandin ajoute à cette définition une différenciation entre chaînes avec répétition de syntagme nominal et chaînes avec démonstratif. Soient les textes

(1) a. *La licorne* à fourrure d'hermine abondait autour du château de X.
b. Un jour *Lancelot* s'amusa à les pourchasser.
c. Piqué au jeu, il les tua toutes.
d. Puis, il les dépouilla et il s'empara de leur précieuse toison.
e. Puis trois jours après, il mourrait dans d'atroces douleurs.
f. *Lancelot* fut pleuré de tous ; Isolde s'enferma dans un couvent...

(2) (a-e) (f) *Lancelot* est le prototype du prédateur aveugle.

(3) (a-e) (f) *La licorne* disparut de la surface de la terre ; elle hante depuis les articles de linguistique et de philosophie.

Ces textes présentent des exemples où un syntagme nominal est répété. L'effet de cette répétition est d'insister sur la nature du référent. En effet, selon Chastain, le contenu descriptif de la deuxième occurrence d'un syntagme nominal répété est différente de la première dans

la mesure où il s'est enrichi des éléments introduits au fur et à mesure du texte. Par exemple, le CD de *Lancelot* en (1f) ou (2f) est /celui qui s'appelle Lancelot, qui a pourchassé, tué, dépouillé les licornes et qui est mort après/. C'est en fonction de ce qu'il a accompli qu'il peut être qualifié de « prototype du prédateur aveugle » en (2f). Dans

(4) (a-e) (f) *Cette vengeance* a frappé de stupeur les contemporains.
Les poètes l'ajoutérent à leur répertoire de faits légendaires.

le démonstratif *cette vengeance* pousse le lecteur à ré-interpréter le texte et notamment l'événement rapporté en (e) : /Lancelot mourir/.

La répétition ou la présence d'un démonstratif associé à un syntagme nominal dans une chaîne-objet font que certains syntagmes nominaux ont la capacité de condenser ce qui est introduit dans cette chaîne. Nous pouvons donc distinguer à l'intérieur d'une chaîne-objet les éléments suivants

- la tête de chaîne constituée d'un syntagme nominal introduisant un objet,
- les segments composés d'éléments qui spécifient l'objet,
- le fermoir constitué d'un syntagme nominal qui a la capacité de condenser la chaîne : *Lancelot* en (1f) et (2f), *la licorne* en (3f) et *cette vengeance* en (4f).

On peut alors faire une différence entre chaînes fermées et chaînes ouvertes, les premières voyant apparaître un syntagme nominal fermoir qui n'est pas mis à jour dans les secondes. Ces deux types de chaînes ont un rapport au thème différent. Dans une chaîne fermée, un syntagme nominal fermoir nomme un objet à partir des spécifications introduites dans la chaîne-objet. L'occurrence du syntagme nominal fermoir a pour effet de transformer ce qui est introduit dans les énoncés où se réalise la chaîne-objet en propriétés caractérisant et spécifiant l'objet qu'il nomme. Les chaînes fermées permettent donc qu'on comprenne les énoncés où elles apparaissent comme étant à propos de quelque chose, ce « quelque chose » étant nommé par le syntagme nominal fermoir. Dans le cas des chaînes ouvertes, le choix du thème n'est pas guidé par cet indice. Comment dès lors nommer le thème de ces énoncés ? Lorsqu'une chaîne-objet ne possède pas de syntagme nominal fermoir, c'est la façon dont est reçu le texte, la lecture de celui-ci, qui va permettre d'inférer le thème. Le lecteur va construire une histoire et un sens qui mettent en jeu l'interprétation des énoncés afin de définir le thème du texte. On passe alors d'une dimension linéaire du thème, abordé grâce aux chaînes fermées, à une dimension synthétique. L'inférence d'un thème peut se déclencher grâce à des transitions temporelles, passage du passé simple à l'imparfait ou du présent au passé par exemple. En effet, lorsqu'il y a une discontinuité dans la présentation de plusieurs événements et que le texte n'explicite pas les raisons de ces discontinuités, alors le travail de lecture du texte peut rechercher, en produisant un thème inféré, à expliciter ce qui rassemble ou disjoint les événements décrits. Cependant cette notion d'inférence, expliquée ici grâce à des indices temporels, ne permet pas de nommer le thème de façon claire et le choix d'un terme lexical n'est pas abordé dans (Marandin, 1988).

À partir du concept de chaînes-objets et après avoir précisé les différences entre chaînes ouvertes et chaînes fermées, Marandin propose cette définition du thème du discours en différenciant thème configuré (à l'aide des syntagmes nominaux fermoirs) et thème inféré (à l'aide de transitions temporelles par exemple).

(i) Un thème est un individu composite, c'est-à-dire un individu relativement à un texte.
(ii) Le contenu descriptif du terme qui le nomme est représentable comme un agrégat subsumant d'autres individus dans leurs interrelations, telles qu'elles sont introduites dans les énoncés, reconstruites dans la compréhension et constitutives d'une interprétation.

Dans le cas d'un thème configuré :

(iii) Le nom du thème est le syntagme nominal en position de fermoir d'une chaîne-objet ;
(iv) Les composants du contenu descriptif sont les segments de la chaîne-objet.

Dans le cas d'un thème inféré :

(iii) Aucune contrainte textuelle ne pèse sur le nom du thème ;
(iv) Les composants du contenu descriptif sont abstraits de la suite d'énoncés.

Dans les deux cas :

(v) La tête nominale du syntagme nominal projette une interprétation sur les composants du contenu descriptif.

Annexe B

Adaptation de l'algorithme de segmentation thématique aux spécificités de documents audiovisuels

Lors de l'adaptation de l'algorithme de segmentation thématique aux spécificités des transcriptions automatiques de vidéos professionnelles, de nombreux paramétrages ont été testés, que ce soit pour l'intégration des mesures de confiance ou lors de la prise en compte d'informations sémantiques ou prosodiques. Nous présentons, dans cette annexe, l'influence des différents paramètres pour chacune des sources d'informations employées. Les tableaux présentés ici contiennent les valeurs de la mesure F_1 obtenues pour une valeur de paramètre α optimum (équation (7.9), page 97), c'est-à-dire pour laquelle le nombre de frontières retournées est le plus proche du nombre de frontières de référence.

B.1 Intégration des mesures de confiance

Dans cette section, nous décrivons les résultats obtenus pour l'intégration des mesures de confiance lors de la segmentation des journaux télévisés (tableau B.1) et des émissions de reportages *Sept à Huit* (tableau B.2).

Dans ces tableaux, la ligne grise correspond aux valeurs calculées lors de l'introduction des mesures de confiance lors de l'estimation du modèle de langue seulement, avec δ_1 variant de 0 à 4, et la colonne grise représente les valeurs fournies par la prise en compte de ces indices uniquement lors du calcul de la probabilité généralisée. La cellule gris foncé correspond, de fait, à la valeur de la mesure F_1 pour une segmentation sans intégration des mesures de confiance (avec $\delta_1 = \delta_2 = 0$; *cf.* section 5.1.1 pour les formules concernées). Dans ces tableaux, nous pouvons observer que l'utilisation des mesures de confiance conduit à une amélioration statistiquement significative (test de Student) de la qualité de la segmentation thématique. En effet, la valeur de la mesure F_1 est augmentée de 2 points pour les journaux télévisés et de 5 points pour les émissions *Sept à Huit*. De plus, ces résultats montrent que l'impact des mesures de confiance est différent pour les deux corpora, selon que ces indices sont intégrés durant l'estimation du modèle de langue ou durant le calcul de la probabilité généralisée. Pour les journaux télévisés, les résultats sont meilleurs lorsque les mesures de confiance sont utilisées pendant le calcul de la probabilité généralisée plutôt que lors de l'estimation du modèle de langue, tandis que pour le corpus d'émissions de reportages, les expériences conduisent à la

TAB. B.1 – Intégration des mesures de confiance pour la segmentation de journaux télévisés

δ_2 | δ_1	0	0,5	1	1,5	2	2,5	3	3,5	4
0	59,7	60,4	60,7	61,0	**61,1**	60,9	60,9	59,7	59,1
0,5	60,4	60,7	61,5	61,5	60,9	60,9	60,8	59,5	58,8
1	61,1	61,5	**61,7**	61,1	60,9	60,9	60,6	58,9	58,8
1,5	61,2	61,6	61,3	60,9	60,5	60,1	60,2	58,9	58,8
2	61,3	61,4	61,4	60,8	60,2	59,9	59,9	58,9	58,7
2,5	61,2	61,5	61,1	60,5	60,1	59,6	59,6	58,6	58,1
3	**61,6**	61,5	60,8	60,4	60,0	59,5	59,4	58,1	57,9
3,5	60,6	60,5	59,3	58,8	58,8	58,3	58,1	57,8	57,8
4	60,9	60,6	59,3	59,0	58,9	58,1	57,8	57,8	57,7

TAB. B.2 – Intégration des mesures de confiance pour la segmentation d'émissions de reportages *Sept à Huit*

δ_2 | δ_1	0	0,5	1	1,5	2	2,5	3	3,5	4
0	54,9	55,0	56,8	54,6	54,7	55,6	58,3	**59,9**	59,1
0,5	53,4	56,1	55,3	56,1	54,7	54,7	55,0	57,9	58,2
1	**55,3**	56,5	55,4	55,6	54,7	55,0	55,0	59,1	58,2
1,5	54,4	55,6	55,6	56,3	54,7	55,0	55,0	56,9	58,6
2	54,9	56,0	56,3	55,0	55,3	55,3	57,0	56,1	56,1
2,5	55,1	55,7	56,3	55,0	55,3	57,0	57,0	56,1	56,5
3	**55,3**	56,8	56,3	55,0	56,7	57,0	58,6	57,7	58,1
3,5	55,3	56,8	56,3	56,3	58,3	58,6	58,6	57,7	57,7
4	54,5	56,8	56,3	**59,9**	59,5	58,6	57,7	57,7	57,1

conclusion inverse. Finalement, une dernière différence existant entre les deux corpora est la valeur optimale de δ_1 et δ_2. Pour le corpus de journaux télévisés, la meilleure valeur de la mesure F_1 est obtenue pour des paramètres δ_1 et δ_2 relativement petits (c'est-à-dire tous deux égaux à 1) tandis que pour les émissions *Sept à Huit* les meilleurs résultats sont fournis grâce à des valeurs de δ_1 et δ_2 plus importantes. Cette différence s'explique par le fait que, pour des grandes valeurs de δ, $c(w_j^i)^\delta$ devient négligeable sauf pour les mots dont la mesure de confiance est très proche de 1. La proportion de mots associés à une mesure de confiance inférieure à 0,9 étant plus importante dans le corpus d'émissions de reportages (36% pour les émissions *Sept à Huit* contre 33% pour les journaux télévisés), l'impact des mesures de confiance est plus perceptible sur ces données, et des valeurs de paramètres plus élevées conduisent à une plus forte amélioration.

B.2 Prise en compte des relations sémantiques

Les tableaux B.3 et B.4, proposés dans cette deuxième section, résument les résultats concernant l'intégration de relations sémantiques lors de la segmentation thématique des corpora de journaux télévisés et d'émissions de reportages respectivement. Dans ces tableaux, les

TAB. B.3 – Prise en compte de relations sémantiques pour la segmentation thématique de journaux télévisés

	Syntagmatique						Paradigmatique					
	ParMot			*Total*			*ParMot*			*Total*		
γ	2	3	10	5k	20k	90k	2	3	10	5k	20k	90k
0,2	59,7	59,7	59,7	59,7	59,7	59,7	59,7	59,7	59,7	59,7	59,7	59,7
0,4	59,7	59,7	60,8	59,7	59,8	59,7	59,7	59,7	59,7	59,7	59,7	60,0
0,5	59,7	59,7	60,1	59,7	60,3	60,4	59,7	59,7	60,0	59,7	59,7	60,3
0,6	59,7	59,6	60,5	59,7	60,0	60,1	59,7	59,7	**60,2**	59,7	60,1	60,1
0,8	**60,0**	59,8	**60,9**	**60,5**	60,6	60,6	59,8	60,3	60,0	59,9	**60,6**	**60,3**
1	58,9	**61,1**	59,8	**60,5**	**60,8**	**61,0**	**60,9**	**60,4**	**60,2**	**60,0**	60,3	**60,3**

TAB. B.4 – Prise en compte de relations sémantiques pour la segmentation thématique d'émissions de reportages *Sept à Huit*

	Syntagmatique						Paradigmatique					
	ParMot			*Total*			*ParMot*			*Total*		
γ	2	3	10	5k	20k	90k	2	3	10	5k	20k	90k
0,2	54,9	54,9	54,9	54,9	54,9	54,9	54,9	54,9	54,9	54,9	54,9	54,9
0,4	54,9	54,9	**55,7**	54,9	**55,5**	**55,5**	54,9	54,9	**55,5**	54,9	54,9	54,9
0,5	54,9	54,9	54,0	54,9	**55,5**	**55,5**	54,9	54,9	55,1	54,9	54,9	**55,1**
0,6	54,9	54,9	53,3	54,9	54,4	54,4	54,9	54,9	54,0	54,9	54,9	54,4
0,8	**56,5**	**55,7**	53,4	**55,5**	53,1	53,1	**55,8**	53,2	53,6	54,9	**55,8**	54,8
1	54,7	54,6	51,6	55,1	54,0	54,0	55,2	53,0	53,0	54,9	52,3	54,0

valeurs pour la prise en compte des relations syntagmatiques sont présentées dans la partie gauche tandis que la partie droite est réservée aux relations paradigmatiques. Pour chaque type de relations sémantiques, les deux méthodes de sélection (*cf.* chapitre 3) ont été employées avec un nombre de relations sémantiques variable : de 2 à 10 pour chaque mot pour la stratégie *ParMot* et de 5 000 à 90 000 pour la technique *Total*. Finalement, les résultats pour la technique de filtrage *Seuil*, utilisée pour ignorer les relations sémantiques des mots entretenant un nombre de relations supérieur à un certain seuil, sont présentés pour une valeur de paramètre γ compris entre $0,2$ et 1.

Pour le corpus de journaux télévisés, la prise en compte de relations sémantiques aide la cohésion lexicale à être plus robuste au faible taux de répétitions de vocabulaire, lié à la taille des segments et à l'utilisation massive de synonymes par les journalistes. En effet, la valeur de la mesure F_1 est augmentée de façon statistiquement significative, de 1,4 point, lors de l'utilisation de relations syntagmatiques et de 1,2 point pour les relations paradigmatiques. Par ailleurs, si les meilleures valeurs de la mesure F_1 sont obtenues grâce aux relations syntagmatiques, l'amélioration globale est plus importante lors de l'utilisation des relations paradigmatiques.

Pour le corpus d'émissions de reportages, l'amélioration proposée par les relations sémantiques est beaucoup plus faible. En effet, un test de significativité a montré que l'utilisation des relations sémantiques sur le corpus *Sept à Huit* ne fournit pas d'amélioration statistiquement

TAB. B.5 – Intégration d'informations prosodiques pour la segmentation thématique de journaux télévisés

	ML				*Proba*				*ML + Proba*			
	MAX	MOY	MIN	ET	MAX	MOY	MIN	ET	MAX	MOY	MIN	ET
intensité	60,67	58,52	57,92	50,27	59,68	58,4	57,58	56,38	59,67	57,15	55,23	40,58
pitch	57,42	57,54	23,63	59,34	59,25	58,11	29,45	58,26	56,82	52,89	**9,73**	54,26
intensité & pitch	56,38	52,4	20,36	44,28	59,81	56,52	27,53	52,75	53,67	40,6	9,81	26,2

significative. Le tableau B.4 montre, par ailleurs, que l'intégration de relations sémantiques dégrade la qualité des résultats lorsque de trop nombreuses relations sont introduites – c'est-à-dire lorsque la valeur du paramètre γ est élevée – dégradation qui est proportionnelle au nombre de relations employé. Ce comportement peut également être observé pour le corpus de journaux télévisés lorsque γ est supérieur à 1. De fait, la technique de filtrage est essentielle pour éviter de prendre en compte des relations sémantiques non adaptées à notre contexte de segmentation thématique, c'est-à-dire des relations liant des mots ou des segments qui ne devraient pas l'être. L'effet plus important de la technique de filtrage sur le corpus d'émissions de reportages peut s'expliquer par le fait que de nombreuses relations sont hors domaine pour ce corpus. De ce fait, des relations sémantiques très générales – associées à des mots tels que « année » ou « aller » – introduisant du bruit, apparaissent plus fréquemment dans ce corpus et ont ainsi plus d'impact sur les émissions de reportages que sur les journaux télévisés.

Finalement, concernant les techniques utilisées pour la sélection des relations sémantiques, aucune différence n'a été observée pour le corpus de journaux télévisés. Au contraire, pour le corpus d'émissions *Sept à Huit*, la méthode de sélection *Total* est meilleure pour les deux types de relations sémantiques. Nous pensons que la méthode *ParMot* sélectionne les relations les plus caractéristiques du corpus sur lequel elles ont été apprises, alors que la technique *Total* est plus susceptible de choisir des relations plus générales. De ce fait, les relations étant extraites d'un corpus d'articles journalistiques, celles sélectionnées grâce à la méthode *ParMot* conviennent davantage au corpus de journaux télévisés et sont moins adaptées à celui d'émissions de reportages.

B.3 Intégration d'informations prosodiques

La dernière section de cette annexe présente les résultats fournis par l'utilisation d'informations prosodiques pour améliorer la qualité de la segmentation thématique de documents audiovisuels. Les tableaux B.5 et B.6 présentent les valeurs de la mesure F_1 obtenues grâce à l'intégration de ces informations lors de la segmentation de journaux télévisés et d'émissions de reportages, respectivement. Les informations prosodiques employées correspondent à deux caractéristiques du signal, l'*intensité* et le *pitch*, pouvant être utilisées ensemble (ligne *intensité & pitch*) ou séparément. Ces informations peuvent être intégrées à l'algorithme lors de l'estimation du modèle de langue – les résultats obtenus dans ce cas sont présentés dans la colonne *ML* – ou du calcul de la probabilité généralisée (colonne *Proba*). La prise en compte de la prosodie peut également s'effectuer lors des deux étapes du calcul de la cohésion lexicale (colonne *ML + Proba*).

Ces tableaux nous montrent, tout d'abord, que l'utilisation d'informations acoustiques

TAB. B.6 – Intégration d'informations prosodiques pour la segmentation thématique d'émissions de reportages *Sept à Huit*

	ML				*Proba*				*ML + Proba*			
	MAX	MOY	MIN	ET	MAX	MOY	MIN	ET	MAX	MOY	MIN	ET
intensité	54,25	57,6	54,82	57,77	52,32	53,6	52,32	53,99	54,53	54,95	51,09	56,91
pitch	56,54	56,07	32,49	58,64	53,96	49,63	35,93	47,44	55,63	49,21	30,63	55,07
intensité & pitch	55,68	50,34	30,9	46,97	53,72	47,44	33,59	51,19	51,71	42,1	**20,3**	47,17

a un impact plus important sur les émissions de reportages que sur le corpus de journaux télévisés, la valeur de la mesure F_1 étant augmentée de 1 point seulement dans le cas des journaux télévisés contre 3,7 points pour les émissions *Sept à Huit*. De plus, nous pouvons remarquer que, pour ces deux corpora, l'impact des informations prosodiques sur la qualité de la segmentation est globalement le même, quelle que soit l'étape de calcul de la cohésion lexicale durant laquelle les informations ont été introduites. Deux éléments se retrouvent cependant dans les deux corpora : les meilleurs résultats (cellules grisées dans les tableaux) sont obtenus lors de la prise en compte des informations prosodiques dans l'estimation du modèle de langue, les plus mauvais étant fournis lors de l'intégration de ces indices lors des deux étapes du calcul de la cohésion lexicale (valeurs en gras).

Finalement, nous remarquons que les plus fortes améliorations, soulignées dans les tableaux, sont obtenues grâce aux techniques d'alignement[1] MAX et ET pour les deux corpora, les deux autres méthodes, MOYENNE et MIN, conduisant à des gains plus faibles ou à des dégradations, ce qui s'explique aisément par le fait que ces deux dernières stratégies favorisent les mots associés aux valeurs de *pitch* ou d'*intensité* les plus faibles.

[1]Une valeur de *pitch* ou d'intensité ayant été extraite pour chaque 0,01 seconde du signal, et la durée de prononciation d'un mot étant généralement supérieure à cette valeur, les quatre stratégies MAX, MOYENNE, MIN et ET consistent à calculer la valeur du score associé au lemme dans la transcription (*cf.* section 3.2 pour plus de détail).

Annexe C

Mise en relation de segments thématiques de programmes TV

Dans cette annexe, nous présentons les résultats fournis par l'adaptation d'un système de mise en relation de segments thématiques à des données télévisuelles (*cf.* chapitre 6 pour les paramètres et les formules cités).

Nous proposons, tout d'abord, les courbes rappel/précision obtenues grâce à la combinaison du critère *tf-idf* et d'informations prosodiques lors de la pondération des vecteurs caractéristiques des segments. Ces courbes nous montrent que, pour le corpus de journaux télévisés, le meilleur paramétrage consiste à associer un poids équivalent aux deux paramètres θ_{ir} et θ_{ac}, quel que soit le type d'informations acoustiques utilisé. Concernant les émissions de reportages *Sept à Huit*, nous pouvons constater sur la figure C.1 que l'influence du critère *tf-idf* doit être plus élevée que celle des informations acoustiques lors du calcul de la pondération des vecteurs. En effet, une importance trop grande donnée à ces dernières dégrade la qualité des résultats.

Nous présentons également, sur la figure C.2, les courbes rappel/précision obtenues lors de l'intégration de relations sémantiques dans le calcul de la similarité entre les différents vecteurs. Nous pouvons remarquer trois éléments importants à partir de ces courbes. Premièrement, les relations paradigmatiques sont plus adaptées à la tâche de mise en relation de segment thématiques que les relations syntagmatiques. En effet, les courbes correspondant à l'intégration des relations paradigmatiques sont au-dessus de celles représentant les syntagmatiques. Deuxièmement, le nombre, ainsi que la qualité des relations introduites, a un impact important sur les résultats. Ainsi, les relations sélectionnées par la méthode *ParMot* conduisent à une meilleure mise en relation que celles sélectionnées grâce à la technique *Total*. Finalement, ces courbes rappel/précision nous montrent que l'utilisation de relations sémantiques ne permet pas d'améliorer la qualité de la mise en relation pour le corpus d'émissions de reportages *Sept à Huit*. Ceci s'explique, selon nous, par le fait que les segments extraits de ces émissions possèdent un vocabulaire plus varié que les journaux télévisés et que l'ajout de relations sémantiques a tendance à introduire du bruit lors de la mise en relation des segments.

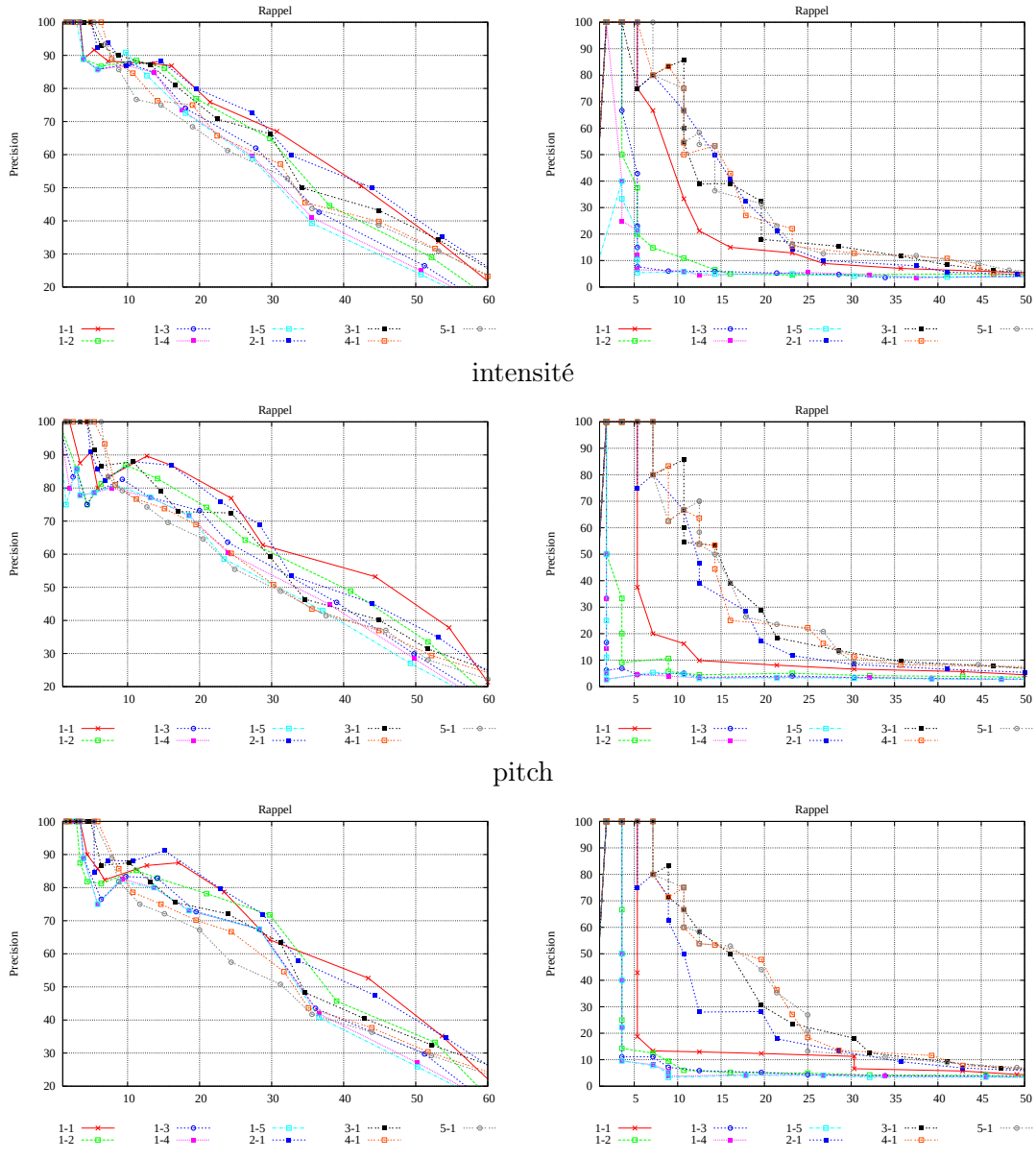


FIG. C.1 – Combinaison du critère *tf-idf* et d'informations acoustiques lors de la pondération des vecteurs caractéristiques pour les corpora de journaux télévisés et d'émissions *Sept à Huit*. Les légendes associées aux courbes correspondent aux valeurs des paramètres θ_{ir}-θ_{ac}

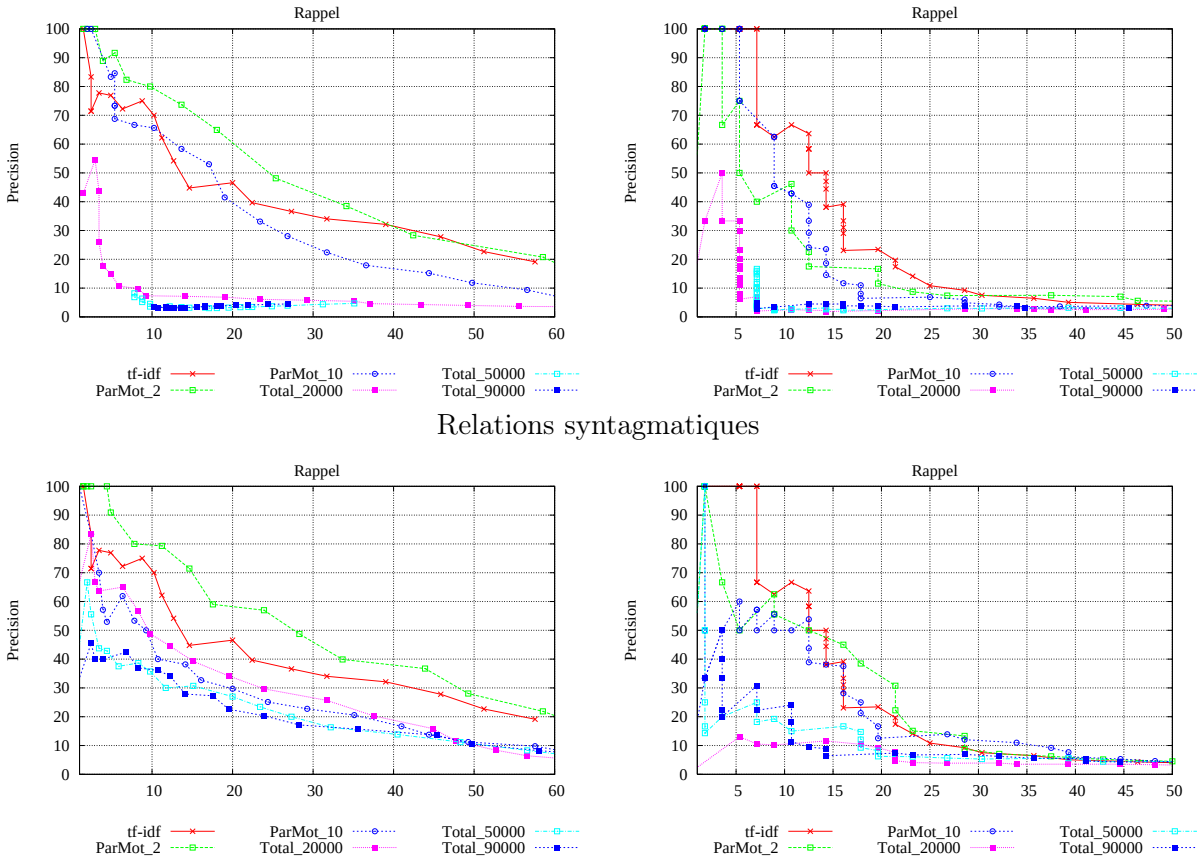


FIG. C.2 – Combinaison du critère *tf-idf* et de relations sémantiques lors du calcul de la similarité entre les segments pour le corpus de journaux télévisés et pour le corpus d'émissions *Sept à Huit*

Annexe D

Impact de la valeur de hiatus Γ lors de l'utilisation de chaînes lexicales pour la segmentation hiérarchique

Dans cette annexe, nous présentons l'impact de la valeur du paramètre Γ, utilisé pour calculer les chaînes lexicales (*cf.* section 7.2.2.1), sur la qualité de la segmentation de segments thématiques. Les figures D.1, D.2 et D.3, proposées dans cette annexe, décrivent les résultats obtenus grâce à des chaînes lexicales calculées à partir de la répétition de lemmes, de *stems* ou prenant en compte des relations sémantiques. Sur ces figures, nous constatons que, pour les trois types de chaînes lexicales employées, une valeur de hiatus Γ égale à 150 fournit les meilleurs résultats. En effet, les courbes roses, représentant la qualité de la segmentation lors de l'intégration de ces chaînes lexicales, sont globalement situées au-dessus des autres courbes.

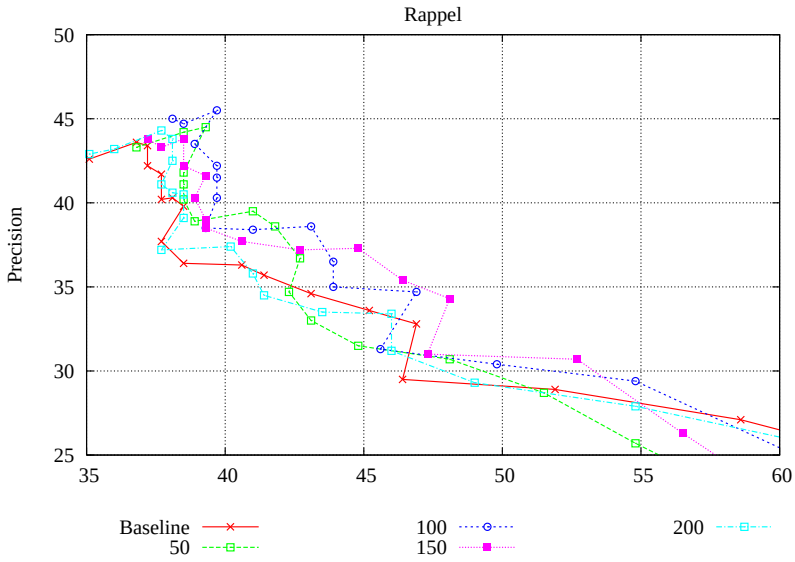


FIG. D.1 – Segmentation thématique obtenue lors de l'intégration de chaînes lexicales reposant sur la répétition de lemmes

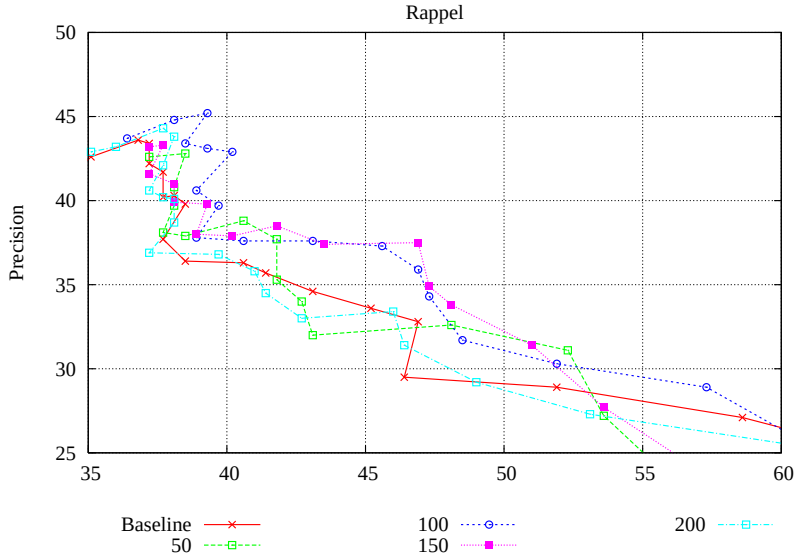


FIG. D.2 – Segmentation thématique obtenue lors de l'intégration de chaînes lexicales reposant sur la répétition de *stems*

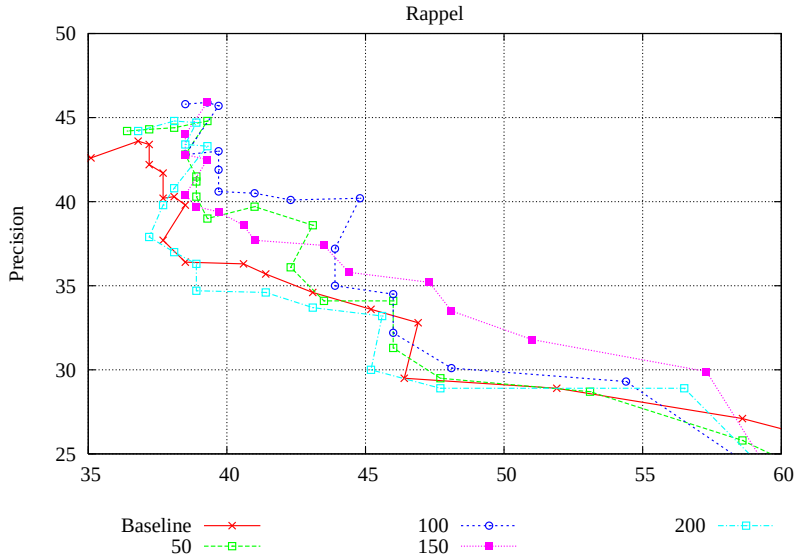


FIG. D.3 – Segmentation thématique obtenue lors de l'intégration de chaînes lexicales prenant en compte des relations sémantiques

Liste de publications

Articles de journaux

1. Guillaume Gravier, Camille Guinaudeau, Gwénolé Lecorvé, Pascale Sébillot. Exploiting speech for automatic TV delinearization : From streams to cross-media semantic navigation. *EURASIP Journal of Image and Video Processing*, 2011(0), 2011.
2. Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot. Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation. *Computer Speech & Language*, 2011, In Press, Corrected Proof.

Conférences internationales

1. Camille Guinaudeau, Julia Hirschberg. Accounting for prosodic information to improve ASR-based topic tracking for TV Broadcast News. In *12th Annual Conference of the International Speech Communication Association, Interspeech'11*, Pages 1401-1404, Florence, Italie, Août 2011.
2. Julien LawTo, Jean-Luc Gauvain, Lori Lamel, Gregory Grefenstette, Guillaume Gravier, Julien Despres, Camille Guinaudeau, Pascale Sébillot. A scalable video search engine based on audio content indexing and topic segmentation. In *4th Annual Conference NEM Summit*, Turin, Italie, Septembre 2011.
3. Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot. Improving ASR-based topic segmentation of TV programs with confidence measures and semantic relations. In *11th Annual Conference of the International Speech Communication Association, Interspeech'10*, Pages 1365-1368, Makuhari, Japon, Septembre 2010.
4. Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot. Utilisation de relations sémantiques pour améliorer la segmentation thématique de documents télévisuels. In *17^{e} conférence sur le traitement automatique des langues naturelles, TALN'10*, Montréal, Québec, Canada, Juillet 2010.
5. Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot. Can Automatic Speech Transcripts be Used for Large Scale TV Stream Description and Structuring ?. In *1st International Workshop on Content-Based Audio/Video Analysis for Novel TV Services, CBTV'09, In conjunction with the International IEEE Symposium on Multimedia, ISM'09*, San Diego, Californie, États-Unis, Décembre 2009.

Conférences nationales

1. Camille Guinaudeau, Guillaume Gravier, Pascale Sébillot. Indices utiles à la cohésion lexicale pour la segmentation thématique de documents oraux. In *28es journées d'étude sur la parole, JEP'10*, Mons, Belgique, Mai 2010.
2. Camille Guinaudeau. Recherche d'information textuelle et phonétique pour le contrôle de l'étiquetage automatique d'émissions dans un flux télévisuel. In *4es rencontres des jeunes chercheurs en recherche d'information, RJCRI'09*, Presqu'île de Giens, France, Mai 2009.

Table des figures

Liste des tableaux

Bibliographie

James Allan. *Topic Detection and Tracking : Event-Based Information Organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002a.

James Allan. Perspectives on information retrieval and speech. In Anni R. Coden, Eric W. Brown, and Savitha Srivinvasen, editors, *Information Retrieval Techniques for Speech Applications*, pages 1–10. Springer, 2002b.

Rui Amaral and Isabel Trancoso. Topic indexing of TV broadcast news programs. In *Proceedings of the International Workshop on Computational Processing of the Portuguese Language*, pages 219 – 226, 2003.

Roxana Angheluta, Rik De Busser, and Marie-Francine Moens. The use of topic segmentation for automatic summarization. In *Proceedings of the Workshop on Text Summarization in Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization*, pages 11–12, 2002.

Michiel Bacchiani and Brian Roark. Unsupervised language model adaptation. In *Proceedings of the 28th International Conference on Acoustics, Speech and Signal Processing*, pages 124 – 127, 2003.

Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL Intelligent Scalable Text Summarization Workshop*, pages 10–17, 1997.

Doug Beeferman, Adam Berger, and John Lafferty. Text segmentation using exponential models. In *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing*, pages 35–46, 1997.

Patrice Bellot and Marc El-Bèze. Classification et segmentation de textes par arbres de décision. *Technique et Science Informatiques*, 20(3) :397 – 424, 2001.

Yves Bestgen. Improving text segmentation using latent semantic analysis : A reanalysis of choi, wiemer-hastings, and moore. *Computational Linguistics*, 32 :455–455, September 2006.

Paul Boersma and David Weenink. Praat, a system for doing phonetics by computer. *Glot International*, 5(9-10) :341 – 345, 2002. http ://www.fon.hum.uva.nl/praat/.

Branimir K. Boguraev and Mary S. Neff. Lexical cohesion, discourse segmentation and document summarization. In *Proceedings of the 6th International Conference on Content-Based Multimedia Information Access*, pages 962 – 979, 2000.

Didier Bourigault. UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9e Conférence sur le Traitement Automatique des Langues Naturelles*, pages 75 – 84, 2002.

Gillian Brown and George Yule. *Discourse analysis*. Cambridge University Press, 1983.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, 18 : 467–479, 1990.

Romain Cabasson and Ajay Divakaran. Automatic extraction of soccer video highlights using a combination of motion and audio features. In *Proceedings of the 11th International Conference on Storage and Retrieval for Image and Video Databases*, pages 272–276, 2003.

Lucien Carroll. Evaluating hierarchical discourse segmentation. In *Proceedings of the 11th International Conference of the North American Chapter of the Association for Computational Linguistics*, pages 993–1001, 2010.

Samuel W.K Chan. Using heterogeneous linguistic knowledge in local coherence identification for information retrieval. *Information Sciences*, pages 313 – 328, 2000.

Charles Chastain. Reference and contexte. *Language Mind and Knowledge*, pages 194 – 269, 1975.

Berlin Chen, Hsin min Wang, and Lin shan Lee. Improved spoken document retrieval by exploring extra acoustic and linguistic cues. In *7th European Conference on Speech Communication Association*, pages 299 – 302, 2001.

Freddy Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the 1st International Conference of the North American Chapter of the Association for Computational Linguistics*, pages 26–33, 2000a.

Freddy Y. Y. Choi. A speech interface for rapid reading. In *Proceedings of IEE colloquium : Speech and Language Processing for Disabled and Elderly People*, pages 1 – 4, 2000b.

Kenneth Church, William Gale, Patrick Hanks, and Donald Hindle. Using statistics in lexical analysis. In U. Zernik, editor, *Lexical Acquisition : Exploiting On-Line Ressources to Build a Lexicon*, pages 115 – 164. Lawrence Erlbaum Associates, 1991.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76 – 83, 1989.

Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 13 :22–29, 1990.

Vincent Claveau and Sébastien Lefèvre. Segmentation thématique : apport de la vectorisation. *Actes de la Conférence francophone en Recherche d'Information et Applications*, 2011a.

Vincent Claveau and Sébastien Lefèvre. Topic segmentation of TV-streams by mathematical morphology and vectorization. In *Proceedings of the 12th International Conference of the International Speech Communication Association*, pages 1105 – 1108, 2011b.

Fabio Crestani. Towards the use of prosodic information for spoken document retrieval. In *Proceedings of the 24th International Conference on Research and Development in Information Retrieval*, pages 420 – 421, 2001.

Béatrice Daille. *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université de Paris 7, 1994.

Béatrice Daille. Conceptual structuring through term variations. In *Proceedings of the ACL 2003 workshop on Multiword expressions : analysis, acquisition and treatment*, pages 9 – 16, 2003.

Gaël de Chalendar and Brigitte Grau. Vers une base de connaissances structurées. In *Actes des Journées internationales d'Orsay sur les Sciences Cognitives*, 2000.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6) :391–407, 1990.

Manolis Delakis. *Multimodal Tennis Video Structure Analysis with Segment Models*. PhD thesis, Université de Rennes 1, France, 2006.

Helge Dyvik. Translations as semantic mirrors : From parallel corpus to WordNet. In *Proceedings of the 23rd International Conference on English Language Research on Computerized Corpora*, pages 311 – 326, 2002.

Stefan Eickeler and Stefan Muller. Content-based video indexing of TV broadcast news using hidden markov models. In *Proceedings of the 24th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2997–3000, 1999.

Jacob Eisenstein. Hierarchical text segmentation from multi-scale lexical cohesion. In *Proceedings of the 10th International Conference of the North American Chapter of the Association for Computational Linguistics*, pages 353–361, 2009.

Benoît Favre, Frédéric Béchet, and Pascal Nocéra. Robust named entity extraction from large spoken archives. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 491–498, 2005.

Julien Fayolle, Fabienne Moreau, Christian Raymond, Guillaume Gravier, and Patrick Gros. CRF-based combination of contextual features to improve a posteriori word-level confidence measures. In *Proceedings of the 11th International Conference on Speech Communication and Technologies*, pages 1942–1945, 2010.

Olivier Ferret. Segmenter et structurer thématiquement des textes par l'utilisation conjointe de collocations et de la récurrence lexicale. In *Actes de la 9e Conférence sur le Traitement Automatique des Langues Naturelles*, pages 155 – 165, 2002.

Olivier Ferret. Approches endogène et exogène pour améliorer la segmentation thématique de documents. *Traitement Automatique des Langues*, 47(2) :111–135, 2006.

Olivier Ferret, Brigitte Grau, and Nicolas Masson. Thematic segmentation of texts : Two methods for two kinds of texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 392–396, 1998.

Stephan Fischer, Rainer Lienhart, and Wolfgang Effelsberg. Automatic recognition of film genres. In *Proceedings of the 3rd International Conference on Multimedia*, pages 295–304, 1995.

Bernard Fradin and Pierre Cadiot. Présentation une crise en thème. *Langue française - volume 78*, pages 3–8, 1988.

Pierre Frath, Rochdi Oueslati, and François Rousselot. Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques. In *Actes des Journées Acquisition Validation et Apprentissage*, 1995.

Fumiyo Fukumoto and Yoshimi Suzuki. Event tracking based on domain dependency. In *Proceedings of the 23rd annual international ACM-SIGIR Conference on Research and Development in Information Retrieval*, 2000.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Proceedings of the 10th International Conference of the International Speech Communication Association*, 2009.

John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees. The TREC spoken document retrieval track : A success story. In *Actes de la 6e Conférence Internationale en Recherche d'Informations Assistée par Ordinateur*, pages 1 – 20, 2000.

Jean-Luc Gauvain, Lori Lamel, and Gilles Adda. The limsi broadcast news transcription system. *Speech Communication*, 37(1-2) :89 – 108, 2002.

Guillaume Gravier, Camille Guinaudeau, Gwénolé Lecorvé, and Pascale Sébillot. Exploiting speech for automatic TV delinearization : From streams to cross-media semantic navigation. *EURASIP Journal of Image and Video Processing*, 2011(0), 2011.

Gregory Grefenstette. Sextant : Exploring unexplored contexts for semantic extraction from syntactic analysis. In *Proceedings of the 30st annual meeting of the Association for Computational Linguistics*, pages 324 – 326, 1992.

Gregory Grefenstette. Corpus-derived first, second and third-order word affinities. In *Proceedings of 6th EURALEX International Congress on Lexicography*, pages 279 – 290, 1994.

Anne Grobet. *L'identification des topiques dans les dialogues.* De Boeck-Duculot, 2002.

Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12 :175–204, 1986.

Camille Guinaudeau and Julia Hirschberg. Accounting for prosodic information to improve asr-based topic tracking for TV broadcast news. In *Proceedings of the 12th International Conference of the International Speech Communication Association*, pages 1401 – 1404, 2011.

Camille Guinaudeau, Guillaume Gravier, and Pascale Sébillot. Can automatic speech transcripts be used for large scale TV stream description and structuring ? In *Proceedings of the 1st International Workshop on Content-Based Audio/Video Analysis for Novel TV Services*, 2009.

Camille Guinaudeau, Guillaume Gravier, and Pascale Sébillot. Enhancing lexical cohesion measure with confidence measures, semantic relations and language model interpolation for multimedia spoken content topic segmentation. *Computer Speech & Language*, In Press, Corrected Proof :–, 2011.

Michael Alexander K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, 1976.

Jean-Paul Haton, Christophe Cerisara, Dominique Fohr, Yves Laprie, and Kamel Smaïli. *Reconnaissance Automatique de la Parole Du signal à son interprétation*. UniverSciences (Paris). Dunod, 2006.

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539 – 545, 1992.

Marti A. Hearst. Multi-paragraph segmentation of expository texts. In *Proceedings of 32nd Annual meeting of the Association for Computational Linguistics*, pages 9 – 16, 1994.

Marti A. Hearst. TextTiling : Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1) :33–64, 1997.

Marti A. Hearst and Christian Plaunt. Subtopic structuring for full-length document access. In *Proceedings of the Special Interest Group on Information Retrieval*, 1993.

Nicolas Hernandez and Brigitte Grau. Analyse thématique du discours : segmentation, structuration, description et représentation. In *Actes du 5e Colloque International sur le Document Électronique*, pages 277 – 285, 2002.

Julia Hirschberg. Communication and Prosody : Functional aspects of Prosody. *Speech Communication*, 36(1-2) :31 – 43, 2002.

Julia Hirshberg and Christine H. Nakatani. Acoustic indicators of topic segmentation. In *Proceedings of the 5th International Conference on Spoken Language Processing*, pages 976 – 979, 1998.

Graeme Hirst and Alexander Budanitsky. Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11 :87–111, 2005.

Winston H. Hsu and Shih-Fu Chang. Topic tracking across broadcast news videos with visual duplicates and semantic concepts. In *Proceedings of the 12th IEEE International Conference on Image Processing*, pages 141 –144, 2006.

Pei-Yun Hsueh, Johanna Moore, and Steve Renals. Automatic segmentation of multiparty dialogue. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 273 – 280, 2006.

Stéphane Huet. *Informations morpho-syntaxiques et adaptation thématique pour améliorer la reconnaissance de la parole*. PhD thesis, Université de Rennes 1, France, 2007.

Stéphane Huet, Guillaume Gravier, and Pascale Sébillot. Un modèle multi-sources pour la segmentation en sujets de journaux radiophoniques. In *Actes de la 15e Conférence sur le Traitement Automatique des Langues Naturelles*, pages 49–58, June 2008.

Stéphane Huet, Guillaume Gravier, and Pascale Sébillot. Morpho-syntactic post-processing of N-best lists for improved French automatic speech recognition. *Computer Speech and Language*, 24 :663–684, 2010.

Ichiro Ide, Koji Yamamoto, Hamada Reiko, and Hidehiko Tanaka. An automatic video indexing method based on shot classification. *Systems and Computers in Japan*, 32 :32 – 41, 2001.

Ichiro Ide, Hiroshi Mo, Norio Katayama, and Shin'ichi Satoh. Topic threading for structuring a large-scale news video archive. In Peter Enser, Yiannis Kompatsiaris, Noel O'Connor, Alan Smeaton, and Arnold Smeulders, editors, *Image and Video Retrieval*, volume 3115 of *Lecture Notes in Computer Science*, pages 2128–2129. Springer Berlin / Heidelberg, 2004.

Ichiro Ide, Tomoyoshi Kinoshita, Hiroshi Mo, Norio Katayama, and Shin'ichi Satoh. *trackThem* : Exploring a large-scale news video archive by tracking human relations. In Gary Lee, Akio Yamada, Helen Meng, and Sung Myaeng, editors, *Information Retrieval Technology*, volume 3689 of *Lecture Notes in Computer Science*, pages 510–515. Springer Berlin / Heidelberg, 2005.

Diana Inkpen and Alain Desilets. Semantic similarity for detecting recognition errors in automatic speech transcripts. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.

Frederick Jelinek. *Statistical Methods for Speech Recognition.* The MIT Press, January 1998.

Frederick Jelinek and Robert L. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397, 1981.

Xiang Ji and Hongyuan Zha. Domain-independent text segmentation using anisotropic diffusion and dynamic programming. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 322–329, 2003.

Sue E. Johnson, Pierre Jourlin, Karen Spärck Jones, and Philip C. Woodland. Spoken document retrieval for TREC-8 at cambridge university. In *Proceedings of the 8th International Conference on Text Retrieval Conference*, pages 197 – 206, 2000.

Athanasios Kehagias, Fragkou Pavlina, and Vassilios Petridis. Linear text segmentation using a dynamic programming algorithm. In *Proceedings of the 10th conference on European chapter of the Association for Computational Linguistics*, pages 171–178, 2003.

Ewa Kijak. *Structuration multimodale des vidéos de sports par modèles stochastiques.* PhD thesis, Université de Rennes 1, 2003.

Beata Beigman Klebanov, Daniel Diermeier, and Eyal Beigman. Lexical cohesion analysis of political speech. *Political Analysis*, 16 :447–463, 2008.

Solomon Kullback and Richard Liebler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1) :79 – 86, 1951.

Alexandre Labadié and Jacques Chauché. Segmentation thématique par calcul de distance thématique. In *Actes des 7e journées francophones Extraction et Gestion des Connaissances*, pages 355–366, 2007.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. Introduction to Latent Semantic Analysis. *Discours Processes*, 25 :259–284, 1998.

Gwénolé Lecorvé, Guillaume Gravier, and Pascale Sébillot. Vers une adaptation thématique non supervisée de modèles de langage - utilisation d'internet comme un corpus ouvert. In *Actes des 27es journées d'étude sur la parole*, pages 149–152, 2008.

Benjamin Lecouteux, Georges Linarès, and Benoit Favre. Combined low level and high level features for out-of-vocabulary word detection. In *Proceedings of the 34th IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2009.

Wen-Syan Li, Necip Fazil Ayan, Okan Kolak, Quoc Vu, Hajime Takano, and Hisashi Shimamura. Constructing multi-granular and topic-focused web site maps. In *Proceedings of the 10th International Conference on World Wide Web*, pages 343–354, 2001.

Liuhong Liang, Hong Lu, Xiangyang Xue, and Yap-Peng Tan. Program segmentation for TV videos. In *Proceedings of the IEEE International Symposium on Circuits and Systems*, pages 1549 – 1552, 2005.

Rainer Lienhart. Reliable transition detection in videos : A survey and practitioner's guide. *International Journal of Image and Graphics*, 1 :469 – 486, 2001.

Diane J. Litman and Rebecca J. Passonneau. Combining multiple knowledge sources for discourse segmentation. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 108–115, 1995.

Zhu Liu, Jincheng Huang, and Yao Wang. Classification TV programs based on audio information using hidden markov model. In *Proceedings of the 2nd IEEE Workshop on Multimedia Signal Processing*, pages 27 –32, 1998.

Beth Logan, Pedro Moreno, and Om Deshmukh. Word and sub-word indexing approaches for reducing the effects of OOV queries on spoken audio. In *Proceedings of the 2nd International Conference on Human Language Technology Research*, pages 31–35, 2002.

Igor Malioutov and Regina Barzilay. Minimum cut model for spoken lecture segmentation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 25–32, 2006.

Okumura Manabu and Honda Takeo. Word sense disambiguation and text segmentation based on lexical cohesion. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 755 – 761, 1994.

William C. Mann and Sandra A. Thompson. Rhetorical structure theory : Toward a functionnal theory of text organisation. *Text*, 3(8) :243 – 281, 1988.

Gaël Manson and Sid-Ahmed Berrani. Automatic TV Broadcast Structuring. *International journal of digital multimedia broadcasting*, 2010, 2010.

Jean-Marie Marandin. A propos de la notion de thème de discours. Éléments d'analyse dans le récit. *Langue française - volume 78*, pages 67–87, 1988.

Daniel McDonald and Hsinchun Chen. Using sentence selection heuristics to rank text segments in txtractor. In *Proceedings of the 2nd ACM/IEEE Joint Conference on Digital Libraries*, pages 28 – 35, 2002.

Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the 2nd international conference on Human Language Technology Research*, pages 280–285, 2002.

David Miller, Richard Schwartz, Ralph Weischedel, and Rebecca Stone. Named entity extraction from broadcast news. In *Proceedings of the DARPA Broadcast News Workshop*, pages 37 – 40, 1999.

George A. Miller. WordNet : a lexical database for English. *Communications of the Association for Computing Machinery*, 38 :39–41, 1995.

Hemant Misra and François Yvon. Modèles thématiques pour la segmentation de documents. In *Actes des 10e Journées Internationales d'Analyse Statistique des données textuelles*, pages 203–213, 2010.

Marie-Francine Moens and Rik De Busser. Generic topic segmentation of document texts. In *Proceedings of the 24th International Conference on Research and Development in Information Retrieval*, pages 418–419, 2001.

Alistair Moffat, Ron Sachs-Davis, Ross Wilkinson, and Justin Zobel. Retrieval of partial documents. In *Proceedings of the 2nd International Conference on Text Retrieval Conference*, pages 181–190, 1994.

Mehryar Mohri, Pedro Moreno, and Eugene Weinstein. A new quality measure for topic segmentation of text and speech. In *Proceedings of the 10th International Conference of the International Speech Communication Association*, pages 2743–2746, 2009.

Mehryar Mohri, Pedro Moreno, and Eugene Weinstein. Discriminative topic segmentation of text and speech. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.

Simon Moncrieff, Chitra Dorai, and Svetha Venkatesch. Detecting indexical signs in film audio for scene interpretation. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 989 – 992, 2001.

Emmanuel Morin. Prométhée, un outil d'aide à l'acquisition de relations sémantiques entre termes. In *Actes de la 5e Conférence sur le Traitement Automatique des Langues Naturelles*, 1998.

Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17 :21–48, 1991.

Paul Van Mulbregt, Ira Carp, Lawrence Gillick, Steve Lowe, and Jon Yamron. Text segmentation and topic tracking on broadcast news via a hidden Markov model approach. In *Proceedings of the 5th International Conference on Spoken Language Processing*, pages 2519 – 2522, 1998.

Philippe Muller and Philippe Langlais. Comparaison de ressources lexicales pour l'extraction de synonymes. In *Actes de la 17e Conférence sur le Traitement Automatique des Langues Naturelles*, 2010.

Jeho Nam, Masoud Alghoniemy, and Ahmed H. Tewfik. Audio-visual content-based violent scene characterization. In *Proceedings of the 4th International Conference on Image Processing*, pages 353 – 357, 1998.

Xavier Naturel. *Structuration automatique de flux vidéos de télévision.* PhD thesis, Université de Rennes 1, 2007.

Stanislas Oger, Mickael Rouvier, and Georges Linarès. Classification du genre vidéo reposant sur des transcriptions automatiques. In *Actes de la 17e Conférence sur le Traitement Automatique des Langues Naturelles*, 2010.

Mari Ostendorf, Benoît Favre, Ralph Grishman, Dilek Hakkani-Tür, Mary Harper, Dustin Hillard, Julia B. Hirschberg, Heng Ji, Jeremy G. Kahn, Yang Liu, Evgeny Matusov, Hermann Ney, Elizabeth Shriberg, Wen Wang, and Chuck Wooters. Speech segmentation and spoken document processing. *IEEE Signal Processing Magazine*, 25(3) :59–69, 2008.

David D. Palmer, Mari Ostendorf, and John D. Burger. Robust information extraction from automatically generated speech transcriptions. *Speech Communication*, 32(1-2) :95 – 109, 2000.

Rebecca J. Passonneau and Diane J. Litman. Intention-based segmentation : Human reliability and correlation with linguistic cues. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 148–155, 1993.

Lev Pevzner and Marti A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28 :19–36, 2002.

Ronan Pichon and Pascale Sébillot. Différencier le sens d'un mot à l'aide du thème et du contexte de leurs occurrences : une expérience. In *Actes de la 6e Conférence sur le Traitement Automatique des Langues Naturelles*, 1999.

Jean-Philippe Poli. *Structuration automatique de flux télévisuels.* PhD thesis, Université Paul Cézanne – Aix-Marseille III, 2007.

Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition.* Prentice Hall, United States edition, 1993.

François Rastier. *Sémantique interprétative.* Presses universitaires de France, 1987.

François Rastier. La sémantique des thèmes - ou le voyage sentimental. In *L'analyse thématique des données textuelles. L'exemple des sentiments*, 1995.

Jeffrey C. Reynar. An automatic method of finding topic boundaries. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, pages 331–333, 1994.

Matthew J. Roach, John S. Mason, and Mark Pawlewski. Video genre classification using dynamics. In *Proceedings of the 26th IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1557 – 1560, 2001.

Mathias Rossignol and Pascale Sébillot. Mise au jour semi-automatique de nuances sémantiques entre mots de sens proches. In *Actes de la 13e Conférence sur le Traitement Automatique des Langues Naturelles*, pages 266 – 275, 2006.

Mathias Rossignol and Pascale Sébillot. Extraction statistique sur corpus de classes de mots-clés thématiques. *Traitement automatique des langues*, 44(3) :217–246, 2003.

Massimo De Santo, Pasquale Foggia, Carlo Sansone, Gennaro Percannella, and Mario Vento. An unsupervised algorithm for anchor shot detection. In *Proceedings of the 18th International Conference on Pattern Recognition*, pages 1238–1241, 2006.

Gerhard Schmid. Treetagger - a language independent part-of-speech tagger. In *Natural Language Processing*, 1994.

Amit Singhal, John Choi, Donald Hindle, David D. Lewis, and Fernando Pereira. AT&T at TREC-7. In *Proceedings of the 7th International Conference on Text Retrieval Conference*, pages 239 – 252, 1999.

Amitabh Kumar Singhal. *Term Weighting Revisited*. PhD thesis, Cornell University, 1997.

Laurianne Sitbon and Patrice Bellot. Évaluation de méthodes de segmentation thématique linéaire non supervisées après adaptation au français. In *Actes de la 11e Conférence sur le Traitement Automatique des Langues Naturelles*, pages 10 – 19, 2004.

Laurianne Sitbon and Patrice Bellot. Segmentation thématique par chaînes lexicales pondérées. In *Actes de la 12e Conférence sur le Traitement Automatique des Langues Naturelles*, pages 505 – 510, 2005.

Gabriel Skantze and Jens Edlund. Early error detection on word level. In *Proceedings of ISCA Tutorial and Research Workshop on Robustness Issues in Conversational Interaction*, 2004.

Malcolm Slaney and Dulce Ponceleon. Hierarchical segmentation : Finding changes in a text signal. In *Proceedings of the 1st International Conference of the Society for Industrial and Applied Mathematics - Text Mining Workshop*, pages 6–13, 2001.

Frank Smadja. Retrieving collocations from text : Xtract. *Computational Linguistics*, 19 : 143–177, 1993.

Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330, 2006.

Alan F. Smeaton, Paul Over, and Aiden R. Doherty. Video shot boundary detection : Seven years of TRECVid activity. *Computer Vision and Image Understanding*, 114(4) :411 – 418, 2010. Special issue on Image and Video Retrieval Evaluation.

Stephen W. Smoliar, Hong J. Zhang, Shuang Y. Tao, and Gong Yihong. Automatic parsing and indexing of news video. *Multimedia Systems*, 2 :256 – 265, 1995.

Nicola Stokes, Joe Carthy, and Alan F. Smeaton. Segmenting broadcast news streams using lexical chains. In *Proceedings of the 1st Starting AI Researchers Symposium*, pages 145 – 154, 2002.

Hui Sun, Guoliang Zhang, Fang Zheng, and Mingxing Xu. Using word confidence measure for OOV words detection in a spontaneous spoken dialog system. In *Proceedings of the 8th European Conference on Speech Communication Association*, pages 2713–2716, 2003.

Gökhan Tür, Dilek Hakkani-Tür, Andreas Stolcke, and Elizabeth Shriberg. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27 : 31– 57, 2001.

Masao Utiyama and Hitoshi Isahara. A statistical model for domain-independent text segmentation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 499–506, 2001.

Michael J. Witbrock and Alexander G. Hauptmann. Using words and phonetic strings for efficient information retrieval from imperfectly transcribed spoken documents. In *Proceedings of the 2nd ACM International Conference on Digital libraries*, pages 30–35, 1997.

Xiao Wu, Chong-Wah Ngo, and Qing Li. Threading and autodocumenting news videos : a promising solution to rapidly browse news topics. *IEEE Signal Processing Magazine*, 23 (2) :59 –68, 2006.

Miao Xingwei. The relationship between cohesion and coherence. *Journal of Foreign Languages*, 1998.

Yaakov Yaari. Segmentation of expository texts by hierarchical agglomerative clustering. In *Proceedings of the 2nd International Conference on the Recent Advances in Natural Language Processing*, 1997.

Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, Thomas Pierce, Brian T. Archibald, and Xin Liu. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems and their Applications*, 14(4) :32 – 43, 1999.

Yiming Yang, Thomas Ault, Thomas Pierce, and Charles W. Lattimer. Improving text categorization methods for event tracking. In *Proceedings of the 23rd annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 65–72, 2000.

Jinhui Yuan, Huiyi Wang, Lan Xiao, Wujie Zheng, Jianmin Li, Fuzong Lin, and Bo Zhang. A formal study of shot boundary detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(2) :168 – 186, 2007.

Zhao Zhao, Shuqiang Jiang, Qingming Huang, and Guangyu Zhu. Highlight summarization in sports video based on replay detection. In *Proceedings of the 4th IEEE International Conference on Multimedia and Expo*, pages 1613 – 1616, 2006.

Résumé

L'augmentation du nombre de documents multimédias disponibles rend nécessaire la mise en place de méthodes de structuration automatique capables de faciliter l'accès à l'information contenue dans les documents, tout en étant suffisamment génériques pour pouvoir structurer des documents tout-venants.

Dans ce cadre, nous proposons deux types de structuration, linéaire et hiérarchique, s'appuyant sur les transcriptions automatiques de la parole prononcée dans les documents. Ces transcriptions, indépendantes du type des documents considérés, sont exploitées par le biais de méthodes issues du traitement automatiques des langues (TAL).

Les deux techniques de structuration, ainsi que la phase de segmentation thématique sur laquelle elles reposent, donnent lieu à plusieurs contributions originales. Tout d'abord, la méthode de segmentation thématique employée, originellement développée pour du texte écrit, est adaptée aux particularités des transcriptions automatiques de vidéos professionnelles – erreurs de transcription, faible nombre de répétitions de vocabulaire. Le critère de cohésion lexicale sur lequel elle se fonde est, en effet, sensible à ces spécificités, ce qui pénalise fortement les performances de l'algorithme. Cette adaptation est mise en place, d'une part grâce à la prise en compte, lors du calcul de la cohésion lexicale, de connaissances linguistiques et d'informations issues de la reconnaissance automatique de la parole et du signal (relations sémantiques, prosodie, mesures de confiance), et d'autre part grâce à des techniques d'interpolation de modèles de langue. À partir de cette étape de segmentation thématique, nous proposons une méthode de structuration thématique linéaire permettant de mettre en relation des segments abordant des thématiques similaires. La méthode employée, fondée sur une technique issue du domaine de la recherche d'information, est adaptée aux données audiovisuelles grâce à des indices prosodiques, qui permettent de favoriser les mots proéminents dans le discours, et des relations sémantiques. Finalement, nous proposons un travail plus exploratoire examinant différentes pistes pour adapter un algorithme de segmentation thématique linéaire à une tâche de segmentation thématique hiérarchique. Pour cela, l'algorithme de segmentation linéaire est modifié – ajustement du calcul de la cohésion lexicale, utilisation de chaines lexicales – pour prendre en compte la distribution du vocabulaire au sein du document à segmenter.

Les expérimentations menées sur trois corpora composés de journaux télévisés et d'émissions de reportages, transcrits manuellement et automatiquement, montrent que les approches proposées conduisent à une amélioration des performances des méthodes de structuration développées.

Abstract

The increasing quantity of video material available requires the implementation of automatic structuring techniques that can facilitate access to the information contained in documents, while being generic enough to be able to structure different kinds of videos.

For this, we develop two kinds of thematic structuring of TV shows, linear or hierarchical, based on the automatic transcripts of the speech pronounced in the programs. These transcripts, independent of the type of documents considered, are used thanks to natural language processing (NLP) methods.

The two structuring techniques, as well as the topic segmentation phase on which they rely, has led to several original contributions. First, the topic segmentation technique employed, originally developed for text, is adapted to the peculiarities of professional videos transcripts – transcription errors, limited number of repetition. The lexical cohesion criterion on which the segmentation step is based is, indeed, sensitive to these characteristics, which severely penalizes the algorithm performances. This adaptation is implemented, on the one hand by taking into account, during the lexical cohesion computation, linguistic knowledge and automatic speech recognition and signal information (semantic relations, prosody, confidence measures), and on the other hand on language model interpolation techniques. From this topic segmentation step, we propose a method for linear thematic structuring that is able to connect segments addressing similar topic. The method, based on a technique from the information retrieval domain, is adapted to the audiovisual data through prosodic cues, that help to promote prominent words in the speech, and semantic relations. Finally, we propose an exploratory work that studies different ways to adapt a linear topic segmentation algorithm to a hierarchical topic segmentation task. For this, the linear topic segmentation algorithm is modified – adjustement of the lexical cohesion computation, use of lexical chains – to reflect the distribution of the vocabulary in the document to be segmented.

Experiments conducted on three corpora composed of broadcast news and reports on current affairs, manually and automatically transcribed, show that the proposed adjustments lead to improved performance of the structuring methods developed.

Utiliser la police Arial Taille 9 dans les champs texte « résumé » et « abstract » - Texte justifié -
Ne pas dépasser le nombre de caractères des cadres de texte ci-dessus.
Ne pas modifier la taille des cadres de texte

N° d'ordre : XXXXXXX
Institut National des Sciences Appliquées de Rennes
20, Avenue des Buttes de Coësmes . CS 14315 . F-35043 Rennes Cedex
Tel : 02 23 23 82 00 – Fax : 02 23 23 83 96